UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

METODOLOGIA BAYESIANA E ADEQUAÇÃO DE MODELOS

Maria João Fernandes Pereira Polidoro

Doutoramento em Estatística e Investigação Operacional

(Especialidade de Probabilidades e Estatística)

2013

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

METODOLOGIA BAYESIANA E ADEQUAÇÃO DE MODELOS

Maria João Fernandes Pereira Polidoro

Tese orientada pela Professora Doutora Maria Antónia Amaral Turkman e pelo Professor Doutor Fernando Magalhães, especialmente elaborada para a obtenção do grau de doutor em Estatística e Investigação Operacional (Especialidade de Probabilidades e Estatística)

2013

Aos meus pais.

Ao meu marido, João Paulo

e aos meus filhos, Sofia e João Pedro.

Resumo

A base de muitas metodologias estatísticas pressupõe que um determinado modelo probabilístico paramétrico se ajusta a um conjunto de dados observados. Se esta suposição falha, a qualidade das inferências realizadas é posta em causa. O estudo da adequabilidade do modelo probabilístico proposto constitui um passo fulcral para o sucesso da modelação estatística.

A literatura estatística clássica que aborda este problema é muito extensa, o mesmo não acontecendo com a literatura bayesiana. Tradicionalmente, a abordagem bayesiana para o estudo da avaliação da adequação de um modelo, compara, através do cálculo de um valor-p preditivo, a amostra observada com amostras simuladas da distribuição preditiva *a posteriori*. No entanto, este método tem sido alvo de muitas críticas. Outros métodos têm sido propostos, em particular testes de ajustamento bayesianos, que requerem a definição de um modelo alternativo ao proposto. A forma de definir o modelo alternativo consiste em incorporar o modelo paramétrico em estudo na família de modelos bayesianos não paramétricos e utilizar, seguidamente, medidas de comparação entre os dois modelos. Destaca-se o factor de Bayes como medida de comparação.

O estudo da adequabilidade de um modelo, seguindo uma abordagem bayesiana não paramétrica, é o tema principal tratado neste trabalho. Efectua-se uma revisão de alguns métodos de estudo da adequação de modelos e propõe-se dois testes bayesianos para o estudo da adequabilidade da distribuição exponencial. São apresentados alguns exemplos práticos para ilustrar alguns dos métodos e é comparado, através de um estudo de simulação, o desempenho dos dois testes com alguns testes de ajustamento clássicos.

Palavras chave: teste de ajustamento bayesiano não paramétrico, teste de ajustamento clássico, factor de Bayes, mistura finita de árvores de Pólya, estudo de simulação, potência do teste.

Abstract

The basis for several statistical methodologies assumes that a specified parametric probabilistic model fits a observed data set. If this assumption does not hold, the quality of the inferences is doubtful. Thus, the study of the adequacy of the proposed probabilistic model is a central issue for the success of statistical modelling.

The classical statistical literature which addresses to this problem is quite wide, in opposition to what happens with Bayesian literature. Traditionally, the Bayesian approach to study and evaluate the adequacy of a model compares the observed sample with simulated samples from the posterior predictive distribution, through the evaluation of a predictive p-value. However, this method has been the subject of much criticism. Other methods have been proposed, particularly Bayesian tests of fit, which require the definition of an alternative model to the proposed one. The way to define the alternative model consists in embedding the parametric model under study in a family of nonparametric Bayesian models and, then, use measures of comparison between the two models. The Bayes factor is one of the most relevant of such measures.

The study of the adequacy of a model, following a nonparametric Bayesian approach, is the main focus of this work. Some methods to study the adequacy of models are presented and two Bayesian tests are proposed aiming to evaluate the adequacy of the exponential model. Some practical examples are presented in order to illustrate the methods and a simulation study is carried out in order to compare the performance of the two methods here proposed with the performance of some classical tests of fit.

Keywords: nonparametric Bayesian test of fit, classical test of fit, finite mixture of Pólya trees, simulation study, power of test.

Agradecimentos

A concretização da presente dissertação não teria sido possível sem o precioso apoio e contributo de algumas pessoas e instituições, às quais aproveito para expressar publicamente o meu agradecimento:

Aos meus orientadores, Professora Doutora Maria Antónia Turkman e Professor Doutor Fernando Magalhães, um agradecimento muito especial, pela orientação científica e sabedoria, pela amizade e pela enorme paciência que sempre tiveram e que foi para mim tão importante.

À Fundação para a Ciência e Tecnologia pelo suporte financeiro, através da bolsa de doutoramento com a referência SFRH/BD/36869/2007 e através do projecto PEst-OE/MAT/UI00006/2011.

Ao Centro de Estatística e Aplicações da Universidade de Lisboa pelo apoio na minha presença em eventos científicos.

À Escola Superior de Tecnologia e Gestão de Felgueiras do Instituto Politécnico do Porto por todo o apoio institucional, bem como ao Instituto Politécnico do Porto pela bolsa que me foi atribuída através do Programa de Formação Avançada de Docentes.

Aos meus pais, irmãos, familiares e amigos pelo incentivo e carinho.

Ao meu marido João Paulo, à minha filha Sofia e ao meu filho João Pedro agradeço todo o amor, carinho e compreensão ao longo desta caminhada.

Índice

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas

AD_n	estatística de Anderson e Darling
BF	factor de Bayes
$\mathrm{BH}_{n,a}$	estatística de Baringhaus e Henze
CPO	ordenada preditiva condicional
CO_n	estatística de Cox e Oakes
$\overline{\mathrm{CM}}_n$	estatística de Cramér-von Mises modificada
DIC	critério de informação da *deviance*
DP	processo de Dirichlet
EP_n	estatística de Epps e Pulley
iid	independentes e identicamente distribuídas
MPT_M	mistura finita de árvores de Pólya com M níveis
MCMC	Monte Carlo via cadeias de Markov
$\mathrm{N}(\mu, \sigma)$	distribuição normal com valor médio μ e desvio padrão σ
Pr	probabilidade
PT	árvore de Pólya
Q_n	estatística do qui-quadrado clássica
Q_n^B	estatística do qui-quadrado bayesiana
$\mathrm{T}_{n,a}$	estatística de Henze e Meintains

Capítulo 1

Introdução

Um dos problemas básicos em modelação estatística é o de averiguar se o modelo probabilístico proposto para representar o fenómeno aleatório que produz um conjunto de dados é ou não adequado.

Considere-se que X é uma variável aleatória (discreta ou contínua) que representa o atributo da população em estudo e que é caracterizada por uma distribuição de probabilidade que pertence a uma família paramétrica $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$, onde apenas se desconhece o verdadeiro valor do vector de parâmetros. Seja $(X_1, X_2, \ldots, X_n)$ uma amostra constituída por n variáveis aleatórias independentes e identicamente distribuídas a X; o problema consiste em testar a hipótese nula $H_0 : X \sim f(x|\theta)$, com base numa amostra observada $(x_1, x_2, \ldots, x_n)$ da amostra aleatória considerada.

O estudo da adequabilidade de um modelo (*model adequacy*), que na literatura estatística encontra vários sinóminos, tais como *model checking*, *model validation*, *model evaluation* ou mesmo *model criticism*, tem ganho popularidade entre a comunidade estatística, em particular desde que Box (1980) identificou o tema como um dos dois principais passos no desenvolvimento da modelação estatística.

Numa primeira fase, pode explorar-se alguns métodos baseados na representação gráfica dos dados observados. Estes métodos dividem-se em: (i) descritivos, tais como

o diagrama de caule-e-folhas (*Stem-and-leaf plot*), o histograma ou o diagrama em caixa (*Boxplot*), onde apenas se pretende realçar as características descritivas dos próprios dados; e (ii) teóricos, onde se pretende verificar como a distribuição dos dados se compara com a distribuição teórica em estudo, como é o caso do gráfico quantil-quantil (*Q-Q plot*). Embora apelativo e útil, o método gráfico não providencia um critério objectivo para concluir sobre a adequabilidade do modelo aos dados. Consequentemente, têm sido desenvolvidos métodos mais formais, na tentativa de quantificar os desvios dos dados em relação à distribuição teórica em estudo.

Na abordagem clássica, o estudo da adequabilidade de um modelo passa pela formulação de um teste de ajustamento (*goodness of fit test*), em que a hipótese nula consiste no modelo proposto e onde não é especificado nenhum modelo alternativo. Há inúmeras estatísticas de teste propostas para este efeito, sendo este um tópico muito investigado. Veja-se, por exemplo, os livros de D'Agostino e Stephens (1986) e Thode (2002) onde os autores apresentam uma descrição detalhada de muitos dos testes de ajustamento clássicos.

A literatura estatística bayesiana sobre métodos para estudar a adequabilidade de um modelo, ao contrário da literatura clássica, é ainda muito reduzida. Desta forma, este tema é um interessante desafio de investigação.

Na abordagem bayesiana, além da especificação de uma distribuição de probabilidade para a variável aleatória que descreve o processo que originou os dados, há que considerar a especificação de uma distribuição *a priori* para o vector de parâmetros que indexa o modelo paramétrico. Este é o modelo bayesiano padrão, definido através de uma hierarquia de dois níveis, traduzindo o primeiro a distribuição dos dados condicional ao vector de parâmetros e o segundo a distribuição *a priori* do vector de parâmetros. No entanto, convém referir que o conceito de modelo, para o estudo da adequabilidade, refere-se apenas à distribuição de probabilidade condicional a um vector de parâmetros desconhecido, ou seja à distribuição amostral.

Os métodos (testes) bayesianos existentes para o estudo da adequabilidade de um

modelo dividem-se em dois tipos de testes; testes globais e testes específicos. Os testes globais são muito semelhantes aos testes de ajustamento clássicos, ou seja, são testes que não requerem a especificação de um modelo alternativo. Consequentemente, as hipóteses nula e alternativa podem ser simplesmente definidas como,

$$H_0 : X \sim f(x|\theta) \; vs \; H_1 : X \nsim f(x|\theta).$$

Para proceder à formulação do teste, define-se uma estatística, $T(X)$, função dos dados, ou uma medida de discrepância, $T(X, \theta)$, função dos dados e do vector de parâmetros, por forma a quantificar o grau de incompatibilidade dos dados com o modelo proposto. Há que distinguir dois métodos genéricos dentro dos testes globais: o método baseado no valor-p (p-*value*) preditivo *a posteriori* proposto por Rubin (1984) e Gelman et al. (1996) e o teste do qui-quadrado de Pearson bayesiano proposto por Johnson (2004). O valor-p preditivo *a posteriori* tornou-se bastante popular e é, actualmente, um dos métodos mais utilizado. No entanto, alguns autores não recomendam o seu uso. As criticas mais frequentes baseiam-se no duplo uso dos dados (Bayarri e Berger, 2000) e na sua calibração (Hjort et al., 2006).

Os testes específicos, ao contrário dos testes globais, requerem a definição de um modelo alternativo (hipótese alternativa, H_1). A forma de definir o modelo alternativo consiste em incorporar (*embed*) o modelo paramétrico em estudo na família de modelos bayesianos não paramétricos. A averiguação da adequabilidade do modelo proposto na hipótese nula é feita através da comparação entre os dois modelos (hipóteses H_0 e H_1). Destaca-se o factor de Bayes (Jeffreys, 1961) como um dos critérios mais utilizados para a comparação de modelos, seguindo uma abordagem bayesiana. O teste específico é vulgarmente designado de teste de ajustamento bayesiano não paramétrico, uma vez que o modelo alternativo é um modelo bayesiano não paramétrico.

O conceito de modelo bayesiano não paramétrico é um tema bastante recente e em grande desenvolvimento, em particular nas áreas da modelação e da estimação (ver, por exemplo, o livro de Hjort et al. (2010)). A literatura estatística bayesiana que aborda o problema de testes bayesianos não paramétricos é ainda muito reduzida

e escassa. Destacam-se os artigos pioneiros de Carota et al. (1996) e Florens et al. (1996). Nesses trabalhos os autores fazem uma primeira abordagem ao tema, limitando o estudo ao caso em que os modelos são discretos, sem deixar de apontar muitas das suas limitações. Ainda dentro do estudo da adequabilidade de modelos discretos, Conigliani et al. (2000) propõem um teste bayesiano não paramétrico onde utilizam o factor de Bayes fraccionário (ver secção 2.3.1), por forma a permitir o uso de distribuições *a priori* não informativas para o vector de parâmetros, em particular distribuições *a priori* impróprias, distribuições estas que são usuais neste tipo de problema.

Para o estudo da adequabilidade de distribuições contínuas, os testes bayesianos não paramétricos encontrados na literatura, dizem respeito ao estudo da adequabilidade do modelo gaussiano ou normal. Verdinelli e Wasserman (1998), Berger e Guglielmi (2001) e Tokdar e Martin (2011) definem como modelo alternativo os seguintes modelos bayesianos não paramétricos: mistura de processos gaussianos, mistura de árvores de Pólya e mistura por processo de Dirichlet, respectivamente. É possível o cálculo do factor de Bayes para todos os testes mas, no caso em que o modelo alternativo é baseado na mistura de árvores de Pólya, o cálculo desse factor é computacionalmente mais acessível. Teoricamente, a sua construção é, também, a mais intuitiva. Além disso, o teste de ajustamento de Berger e Guglielmi (2001) pode ser aplicado a outros modelos. Ou seja, é um teste de ajustamento não paramétrico que se pode construir e definir para o estudo da adequabilidade de outras distribuições, para além da distribuição normal.

O estudo da adequabilidade de um modelo seguindo uma abordagem bayesiana não paramétrica é o tema principal deste trabalho. O modelo exponencial tem particular interesse, por exemplo, na Teoria da Fiabilidade para modelar tempos de vida, sendo talvez o modelo mais usado nessa área. Na abordagem clássica há muitos testes propostos para testar quer a normalidade quer a exponencialidade de uma amostra univariada com vector de parâmetros desconhecido. A comparação da qualidade dos diferentes testes em termos de potência é, em geral, feita através de estudos de simulação. Veja-se, por exemplo, os trabalhos recentes de Baringhaus e Henze (2000), Choi et al. (2004),

Henze e Meintanis (2005) e Grané e Fortiana (2011) para o estudo da adequabilidade da distribuição exponencial e os trabalhos de Zhang e Wu (2005), Coin (2008) e Romão et al. (2010) para o estudo da adequabilidade da distribuição normal, assim como muitas outras referências relevantes nesses artigos.

Um dos objectivos deste trabalho é propor um teste bayesiano não paramétrico para o estudo da adequabilidade da distribuição exponencial, que comparativamente com os testes de ajustamento clássicos existentes, tenha um desempenho melhor, em termos de potência.

No Capítulo 2, são apresentados conceitos fundamentais e os modelos bayesianos em estudo, ou seja, o modelo paramétrico e o modelo não paramétrico. De entre os modelos bayesianos não paramétricos, é dado especial destaque à distribuição *a priori* mistura finita de árvores de Pólya, uma vez que esta é de primordial importância para a definição do teste de ajustamento bayesiano não paramétrico a desenvolver. No terceiro Capítulo, apresenta-se uma revisão de alguns dos métodos para o estudo da adequabilidade de um modelo. No Capítulo 4, são propostos dois testes bayesianos para o estudo da adequabilidade da distribuição exponencial. São ainda apresentados alguns exemplos práticos para ilustração dos métodos e é comparado, através de um estudo de simulação Monte Carlo, o desempenho dos dois testes bayesianos com alguns testes de ajustamento clássicos. No Capítulo 5 é feito um resumo das principais contribuições deste trabalho, assim como das conclusões obtidas.

A notação e abreviaturas utilizadas neste trabalho, são apresentadas na primeira vez que aparecem em cada um dos capítulos. Todos os gráficos e trabalho computacional realizado, foi efectuado utilizando a linguagem de programação R (R Development Core Team, 2011).

O anexo A contém o código em linguagem R para a implementação do algoritmo 4 apresentado no Capítulo 4, referente ao estudo da adequabilidade da distribuição exponencial a um conjunto de dados. O código encontra-se também disponível para ser utilizado pelo leitor no seguinte endereço: https://sites.google.com/site/polidoromjcodigor/.

Capítulo 2

Conceitos fundamentais

2.1 O modelo bayesiano paramétrico

Os problemas estatísticos são descritos através de modelos probabilísticos. A análise estatística de um certo fenómeno aleatório parte de um determinado conjunto de dados observados (ou amostra), seja $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, e pressupõe que este conjunto de dados é a realização de um vector aleatório $X = (X_1, X_2, \ldots, X_n)$. Designa-se por espaço amostral, e representa-se por Ω, o respectivo contradomínio do vector aleatório X, ou seja, o conjunto dos possíveis valores que a amostra pode tomar.

Seguidamente, define-se uma distribuição de probabilidade conjunta, seja F, que representa a incerteza ou variabilidade na observação de X. A forma da distribuição de probabilidade conjunta F não é, obviamente, totalmente conhecida. No entanto, o estatístico possui algum conhecimento inicial sobre o processo gerador do conjunto de dados, levando-o a definir uma família de distribuições de probabilidade $\mathcal{F}$, à qual pertence F, e que se designa por modelo estatístico para X (Paulino et al., 2003).

A família de distribuições $\mathcal{F}$, caracterizada normalmente por funções de distribuição, funções de probabilidade, funções densidade de probabilidade, ou no caso geral, por medidas de probabilidade, está indexada por um vector de parâmetros de dimensão

finita, $\theta = (\theta_1, \theta_2, \ldots, \theta_s) \in \Theta \subseteq \mathbb{R}^s$, com domínio num conjunto Θ, que se designa por espaço paramétrico. O modelo estatístico para X, por exemplo, pode escrever-se

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta,\ x \in \Omega\}.$$

Num modelo bayesiano, além da especificação de uma distribuição de probabilidade conjunta para o vector de dados condicional a um vector de parâmetros, $f(x|\theta)$, há que considerar a especificação de uma distribuição de probabilidade para o próprio vector de parâmetros. Esta distribuição é denominada por distribuição *a priori* de θ, representada por $h(\theta)$, e assenta numa atribuição, usualmente subjectiva, de graus de credibilidade aos diferentes valores de θ, isto é, expressa de alguma forma o conhecimento prévio à observação do conjunto de dados, do estatístico ou investigador, sobre os possíveis valores de θ. Quando em determinado estudo o investigador tem pouca ou nenhuma informação *a priori*, usualmente conhecida como "ignorância *a priori*", considera-se uma distribuição *a priori* não-informativa como, por exemplo, a distribuição *a priori* de Jeffreys (1961).

Informação mais detalhada acerca da representação da informação *a priori* pode ser encontrada em Paulino et al. (2003, Capítulo 2) e nas referências aí indicadas.

Toda a inferência bayesiana sobre o vector de parâmetros do modelo é baseada na actualização da informação inicial de que se dispõe sobre os parâmetros, após se observar o conjunto de dados x. O teorema de Bayes é a regra de actualização utilizada para quantificar essa passagem, e a distribuição resultante é designada por distribuição *a posteriori* de θ, representada por $h(\theta|x)$.

Suponha-se que X_i, para $i = 1, 2, \ldots, n$, são variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) condicionalmente a θ; tem-se então que

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

onde $f(x_i|\theta)$ é a distribuição da variável aleatória X_i condicional a θ.

Considere-se que foi observada a amostra x. A distribuição de probabilidade conjunta para X e θ é dada por

$$f(x, \theta) = f(x|\theta)h(\theta),$$

onde $f(x|\theta)$ é também denominada de informação amostral.

A informação amostral é uma função de duas componentes, x e θ. Fixando θ, $f(\cdot|\theta)$ é uma distribuição de probabilidade. Por outro lado, após observar $X = x$, $f(x|\theta)$ é apenas uma função de θ e, neste caso, passa a ser denominada por função de verosimilhança de θ com respeito ao conjunto de dados observados x, representada também por $L(\theta|x) = f(x|\theta)$. Desta forma, a função de verosimilhança desempenha um papel importante na determinação da distribuição *a posteriori*, pois é interpretada como um meio através do qual o conjunto de dados transforma o conhecimento *a priori* sobre θ (Paulino et al., 2003).

Finalmente, condicional aos dados observados x, a distribuição de probabilidade de θ, ou seja a distribuição *a posteriori* de θ, é dada por

$$h(\theta|x) = \frac{f(x, \theta)}{p(x)} = \frac{f(x|\theta)h(\theta)}{\int_\Theta f(x, \theta)d\theta} = \frac{f(x|\theta)h(\theta)}{\int_\Theta f(x|\theta)h(\theta)d\theta} \propto f(x|\theta)h(\theta), \tag{2.1}$$

onde $p(x)$ define a distribuição marginal de X, também designada por distribuição preditiva *a priori*, uma vez que sumaria a informação relativa a x antes de este ter sido observado. Note-se que, ao omitir o termo $p(x)$, a igualdade definida em (2.1) foi substituída por uma proporcionalidade. Esta forma simplificada será útil em problemas que envolvam estimação de parâmetros, uma vez que o denominador é apenas uma constante normalizadora. No entanto, noutras situações, tem um papel importante como, por exemplo, na comparação de modelos.

Depois do conjunto de dados x ter sido observado, se se quiser fazer predições sobre um conjunto de dados, seja $y = (y_1, y_2, \ldots, y_m)$, desconhecido mas observável, de um modelo cuja distribuição amostral de Y pode ser igual ou diferente da distribuição amostral de X, mas em que ambos os modelos partilham do mesmo vector de parâmetros

θ, define-se a distribuição preditiva *a posteriori*, $p(y|x)$, que é dada por

$$p(y|x) = \int_{\Theta} p(y, \theta|x) d\theta = \int_{\Theta} f(y|x, \theta) h(\theta|x) d\theta.$$

Esta distribuição resume a informação relativa à plausibilidade de observar um conjunto de dados y, condicional ao conjunto de dados observados x. Se, condicionalmente a θ, as observações a predizer forem independentes de X, então $f(y|x, \theta) = f(y|\theta)$ e

$$p(y|x) = \int_{\Theta} f(y|\theta) h(\theta|x) d\theta.$$

Uma das principais vantagens da estatística bayesiana reside no facto de a interpretação probabilística das suas inferências ser mais intuitiva do que na estatística clássica. As principais desvantagens são a dificuldade de atribuir uma distribuição *a priori* para o vector de parâmetros e o cálculo, na maioria das situações complexo, dos integrais necessários à realização de inferências. Para um estudo mais aprofundado dos fundamentos da inferência bayesiana e das metodologias existentes para definir a informação *a priori*, veja-se Paulino et al. (2003) e as referências aí mencionadas. Quanto ao problema do cálculo dos integrais, destaca-se a classe de algoritmos denominada de Monte Carlo via cadeias de Markov (MCMC, *Markov chain Monte Carlo*), em particular o método de amostragem Gibbs (*Gibbs Sampler*). Para mais detalhes sugere-se Paulino et al. (2003, Capítulo 8) e Gamerman e Lopes (2006).

2.2 O modelo bayesiano não paramétrico

O modelo bayesiano é designado de paramétrico se a distribuição de probabilidade utilizada para modelar os dados, $F \in \mathcal{F}$, tem uma forma conhecida e está indexada por um vector de parâmetros de dimensão finita, usualmente desconhecido, mas especificado *a priori*. Simbolicamente, no caso em que se tem uma amostra aleatória simples

$$\begin{array}{c} X_1, X_2, \ldots, X_n|\theta \overset{\text{iid}}{\sim} f(x|\theta) \\ \theta \sim h(\theta) \end{array}.$$

No entanto, conforme refere Jara (2008), "*In many situations, however, constraining inference to a specific parametric form may limit the scope and type of inference that can be drawn from such models. Therefore, we would like to relax parametric assumptions to allow greater modeling flexibility and robustness against mis-specification of a parametric statistical model. In the Bayesian context such flexible inference is typically achieved by placing a prior distribution on infinite-dimensional spaces, such as the space of all probability distributions for a random variable of interest. These models are usually referred to as Bayesian nonparametric models.*"

Ou seja, na abordagem bayesiana não paramétrica, a maneira de flexibilizar a forma da distribuição de probabilidade que modela os dados, por exemplo, dada por G[1], $\{G : G \in \mathcal{G}\}$, consiste em: (i) definir um espaço de medidas de probabilidade sobre um espaço mensurável $(\Omega, \mathcal{B})$, onde Ω é um espaço amostral e $\mathcal{B}$ é uma σ-álgebra de subconjuntos de Ω e, (ii) utilizar medidas de probabilidade aleatórias (*Random Probability Measures*), que são distribuições de probabilidade sobre $\mathcal{G}$, o espaço de todas as medidas de probabilidade. Neste sentido, seja G uma distribuição de probabilidade sobre $(\Omega, \mathcal{B})$, desconhecida; a abordagem bayesiana não paramétrica pressupõe que G é um parâmetro desconhecido e como tal poder atribuir-se-lhe uma distribuição de probabilidade *a priori*, denominada medida de probabilidade aleatória e representada por $\mathcal{P}$. Simbolicamente,

$$\begin{array}{c} X_1, X_2, \ldots, X_n | G \overset{\text{iid}}{\sim} G \\ G \sim \mathcal{P} \end{array} .$$

O espaço amostral Ω ou é discreto, finito ou infinito numerável, ou é contínuo (neste caso, por exemplo $\Omega = \mathbb{R}$). A questão relevante, para a qual é importante ter uma resposta, é: "É possível construir informação *a priori* subjectiva num espaço de dimensão infinita?" (Ghosh e Ramamoorthi, 2003).

Antoniak (1974) refere que se deve tomar em consideração as seguintes propriedades,

[1]Substitui-se F por G para distinguir os dois modelos e porque vão ser utilizados, simultaneamente, mais à frente.

quando se pretende definir uma distribuição *a priori*:

I) A família de distribuições *a priori* deve ser analiticamente tratável nos seguintes aspectos:

   (a) ser simples a determinação da distribuição *a posteriori*,

   (b) ser possível calcular os valores esperados de funções de perda simples,

   (c) ser fechada, ou seja, a distribuição *a posteriori* deve pertencer à mesma família da distribuição *a priori*;

II) A família de distribuições *a priori* deve ser capaz de expressar qualquer informação ou conhecimento sobre o vector de parâmetros;

III) A família de distribuições *a priori* deve ser parametrizada de forma a produzir uma interpretação clara das crenças *a priori*.

Infelizmente, estas propriedades não são satisfeitas simultaneamente. Usualmente, o que acontece é sacrificar alguma(s) para satisfazer outra(s). Ferguson (1973) introduziu o processo de Dirichlet (*Dirichlet process*) como uma medida de probabilidade aleatória que satisfaz a primeira e a terceira propriedade mas é ligeiramente ineficiente em relação à segunda propriedade. No seu trabalho, discute as propriedades básicas da distribuição processo de Dirichlet, prova a sua existência e apresenta uma variedade de problemas de estimação não paramétrica, proporcionando assim uma primeira interpretação bayesiana para alguns dos procedimentos não paramétricos mais comuns.

Uma vantagem importante da utilização de métodos bayesianos não paramétricos, relativamente aos métodos bayesianos paramétricos, consiste na capacidade de incorporar a incerteza ao nível das funções de distribuição (Jara, 2008). No entanto, esta flexibilidade aumenta a complexidade computacional da análise. Por conseguinte, o desenvolvimento dos modelos bayesianos não paramétricos, nas últimas duas décadas, deve-se, em parte, ao grande desenvolvimento das metodologias computacionais utilizando métodos MCMC. Escobar (1988, 1994) foi o primeiro investigador a implementar

a distribuição processo de Dirichlet, utilizando métodos MCMC, em particular o método de amostragem Gibbs, abrindo assim o caminho para contornar os problemas computacionais da inferência bayesiana não paramétrica.

Seguidamente, apresentam-se alguns conceitos básicos da modelação não paramétrica, indicando algumas referências importantes, com o objectivo de enquadrar o problema do estudo da adequabilidade de um modelo, tema do trabalho a desenvolver.

2.2.1 Conceitos básicos

Na primeira parte desta secção introduz-se a notação e apresentam-se as propriedades de algumas distribuições, para melhor compreender a abordagem não paramétrica, em particular a determinação da medida de probabilidade aleatória $\mathcal{P}$. Na literatura estatística não paramétrica encontram-se várias medidas de probabilidade aleatórias, sendo a distribuição processo de Dirichlet a mais referenciada e estudada. Inicia-se esta apresentação das medidas de probabilidade aleatórias com a descrição da distribuição processo de Dirichlet (Ferguson, 1973), apresentando as suas limitações e modelos alternativos. Seguem-se as distribuições árvores de Pólya (*Pólya trees*) que são uma generalização da distribuição processo de Dirichlet (Lavine, 1992, 1994) e finaliza-se com a distribuição mistura de árvores de Pólya (*Mixture Pólya Trees*) (Lavine, 1992; Hanson e Johnson, 2002). Outras medidas de probabilidade aleatórias e alguns métodos de inferência bayesiana não paramétrica podem ser vistos em Ghosh e Ramamoorthi (2003), Hanson (2006) e em Hjort et al. (2010), bem como nas referências aí mencionadas.

Para motivar a ideia geral subjacente ao processo Dirichlet, começa-se por considerar um problema paramétrico simples (ver Wakefield e Walker, 1997).

A distribuição beta

Suponha-se que o espaço amostral discreto Ω é constituído apenas por dois valores distintos e suponha-se que X_i é uma variável aleatória que toma um dos dois valores distintos de Ω, com probabilidades p_1 e p_2.

A incerteza acerca da distribuição de probabilidade desconhecida G é equivalente à incerteza acerca dos valores para (p_1, p_2) ou, simplesmente, para p_1 uma vez que $p_1, p_2 \geq 0$ e $p_1 + p_2 = 1$ e, por isso, $p_2 = 1 - p_1$. Um estatístico bayesiano modela esta incerteza atribuindo uma distribuição de probabilidade *a priori* para as probabilidades desconhecidas.

Como $0 < p_1 < 1$, qualquer distribuição de probabilidade no intervalo $]0, 1[$ define uma distribuição de probabilidade *a priori* para p_1. Em particular, uma distribuição bastante flexível é a distribuição beta, com função densidade de probabilidade

$$h(p_1|\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p_1^{\alpha_1 - 1}(1 - p_1)^{\alpha_2 - 1}, \quad \alpha_1, \alpha_2 > 0$$

onde (α_1, α_2) são os parâmetros da distribuição beta e $\Gamma(\cdot)$ representa a função gama, definida por $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$, $a > 0$. Simbolicamente, $p_1|\alpha_1, \alpha_2 \sim \text{Be}(\alpha_1, \alpha_2)$. Como (p_1, p_2) define uma medida de probabilidade sobre Ω, então a distribuição beta pode ser vista como uma distribuição de probabilidade sobre distribuições de probabilidades.

A atribuição dos valores para os parâmetros (α_1, α_2) é feita de modo a traduzir diferentes opiniões *a priori* para (p_1, p_2). Suponha-se que $\alpha_j = cq_j$, com $c > 0$, $q_j \geq 0$ e $q_1 + q_2 = 1$. Denomine-se as crenças *a priori* (q_1, q_2) por G_0. Então, pelas propriedades da distribuição beta, tem-se que

$$E[p_j] = \frac{\alpha_j}{\alpha_1 + \alpha_2} = q_j$$

e

$$Var[p_j] = \frac{q_j(1 - q_j)}{c + 1}.$$

Desta forma, q_j é entendida como uma medida de probabilidade que centra as crenças *a priori* e c reflecte o grau de certeza nessas crenças. Um valor de c grande implica uma variância pequena e, portanto, uma forte crença *a priori*. É usual denominar c por parâmetro de concentração (ou precisão). Por exemplo, suponha-se que se define que $q_1 = 0.5$ e $c = 2$, então, tem-se que $p_1 \sim \text{Be}(1, 1)$, que resulta na atribuição de uma distribuição *a priori* uniforme para p_1.

Na Figura 2.1, apresentam-se 4 diagramas em caixa, associados a 100 valores simulados da distribuição beta, com $q_1 = 0.5$, isto é, $p_1 \sim \text{Be}(\alpha_1 = cq_1, \alpha_2 = c(1-q_1))$, e para 4 diferentes valores de c. Quando $c = 2$ os 100 valores simulados provêm de uma distribuição $\text{Be}(1,1)$ e daí os valores estarem distribuídos ao longo do intervalo $]0,1[$. À medida que c aumenta, os 100 valores simulados estão mais concentrados em torno de $q_1 = 0.5$.

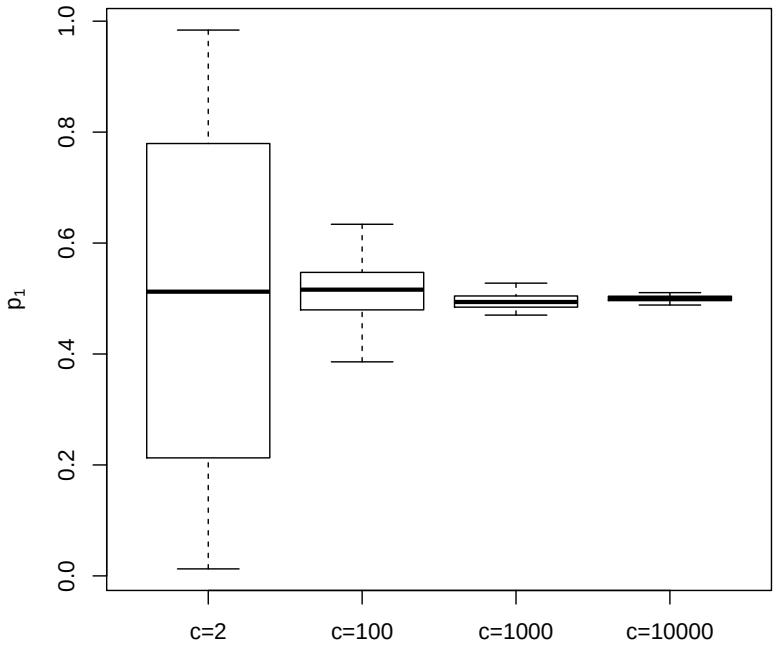

Figura 2.1: Diagramas em caixa de 100 valores simulados da distribuição beta, $\text{Be}(0.5c, 0.5c)$, para c =2, 100, 1000 e 10000.

A escolha da distribuição beta para distribuição *a priori* para p_1 é útil uma vez que ela é a distribuição conjugada natural da distribuição binomial. Desta forma, se $x = (x_1, x_2, \ldots, x_n)$ é a concretização de n variáveis aleatórias independentes e identicamente distribuídas a G, então a distribuição *a posteriori* de (p_1, p_2) é ainda uma distribuição beta, $\text{Be}(\alpha_1+n_1, \alpha_2+n_2)$, onde n_j, para $j = 1, 2$, é o número de observações, na amostra,

iguais a cada um dos dois valores possíveis do espaço amostral Ω, com $n_1 + n_2 = n$.

A distribuição Dirichlet

Considere-se agora que, o espaço amostral discreto e finito Ω é constituído por k valores distintos e suponha-se que X_i é uma variável aleatória que toma qualquer um dos k valores de Ω, cada um com probabilidade p_j, para $j = 1, 2, \ldots, k$, respectivamente, sujeitos a $p_j \geq 0$ e $\sum_{j=1}^{k} p_j = 1$. A incerteza acerca da distribuição de probabilidade desconhecida G é, agora, equivalente à incerteza acerca dos valores para $(p_1, p_2, \ldots, p_k)$ ou, simplesmente, para $(p_1, p_2, \ldots, p_{k-1})$, uma vez que $p_k = 1 - \sum_{j=1}^{k-1} p_j$.

A distribuição de probabilidade *a priori* natural para $(p_1, p_2, \ldots, p_{k-1})$ é, agora, definida pela distribuição Dirichlet com parâmetros $(\alpha_1, \alpha_2, \ldots, \alpha_k) \in \mathbb{R}_+^k$, que generaliza a distribuição beta com parâmetros (α_1, α_2), atrás apresentada. E, neste caso, a distribuição Dirichlet é a distribuição *a priori* conjugada natural da distribuição multinomial (esta distribuição é a extensão multivariada da distribuição binomial).

Considere-se $p = (p_1, p_2, \ldots, p_{k-1})$ e $\alpha = (\alpha_1, \ldots, \alpha_k)$. Então, a função densidade da distribuição Dirichlet é expressa por

$$h(p|\alpha) = \frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \cdots (1 - \sum_{j=1}^{k-1} p_j)^{\alpha_k - 1};$$

simbolicamente, $p|\alpha \sim \mathrm{D}_{k-1}(\alpha)$.

De modo análogo ao caso da distribuição beta, atrás apresentado, a atribuição dos valores para os parâmetros da distribuição Dirichlet, $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, é feita de modo a traduzir diferentes opiniões *a priori* para $(p_1, p_2, \ldots, p_k)$. Supondo que $\alpha_j = cq_j$, agora para $j = 1, 2, \ldots, k$, obtém-se a mesma interpretação apresentada anteriormente, isto é, $(q_1, q_2, \ldots, q_k)$ representa as crenças *a priori* (G_0) e c a certeza nessas crenças.

Agora, se $x = (x_1, x_2, \ldots, x_n)$ é a concretização de n variáveis aleatórias independentes e identicamente distribuídas a G, a distribuição *a posteriori* de p é ainda uma distribuição Dirichlet, $\mathrm{D}_{k-1}(\alpha_1 + n_1, \alpha_2 + n_2, \ldots, \alpha_n + n_k)$, onde n_j, para $j = 1, 2, \ldots, k$, é o número de observações, na amostra, iguais a cada um dos k valores possíveis do

espaço amostral Ω, com $n_1 + n_2 + \cdots + n_k = n$.

A distribuição Dirichlet goza de muitas propriedades importantes. Seguidamente apresentam-se duas delas.

A distribuição marginal de cada p_j, para $j = 1, 2, \ldots, k$, segue uma distribuição beta; simbolicamente

$$p_j \sim \mathrm{Be}(\alpha_j, \sum_{i=1}^{k} \alpha_i - \alpha_j).$$

Se $(p_1, p_2, \ldots, p_k) \sim \mathrm{D}_{k-1}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ e se se formarem agrupamentos B_l, para $l = 1, 2, \ldots, m$ dos p_j, para $j = 1, 2, \ldots, k$, $m < k$, então prova-se que a distribuição resultante ainda é uma distribuição Dirichlet com parâmetros iguais à soma dos parâmetros correspondentes aos p_j agrupados, isto é,

$$\begin{gathered}\left(\textstyle\sum_{j\in B_1} p_j, \sum_{j\in B_2} p_j, \ldots, \sum_{j\in B_m} p_j\right) \\ \sim \mathrm{D}_{m-1}\left(\textstyle\sum_{j\in B_1} \alpha_j, \sum_{j\in B_2} \alpha_j, \ldots, \sum_{j\in B_m} \alpha_j\right).\end{gathered}$$

Por outras palavras, para uma qualquer partição[2] $(B_1, B_2, \ldots, B_m)$ de Ω, o vector das probabilidades sobre a partição

$$(\Pr(B_1), \Pr(B_2), \ldots, \Pr(B_m)) \sim \mathrm{D}_{m-1}(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_m)),$$

onde $\Pr(B_l) = \sum_{j\in B_l} p_j$ e $\alpha(B_l) = \sum_{j\in B_l} \alpha_j$, para $l = 1, 2, \ldots, m$ e $j = 1, 2, \ldots, k$.

Agora, fazendo $\alpha(B_l) = cG_0(B_l)$, para $l = 1, 2, \ldots, m$, pode então escrever-se simplesmente

$$\begin{gathered}(\Pr(B_1), \Pr(B_2), \ldots, \Pr(B_m)) \sim \\ \mathrm{D}_{m-1}(cG_0(B_1), cG_0(B_2), \ldots, cG_0(B_m)).\end{gathered}$$

[2]Uma partição $(B_1, B_2, \ldots, B_m)$ do espaço amostral Ω é tal que $\bigcup_{l=1}^{m} B_l = \Omega$ e $B_l \cap B_{l^*} = \emptyset$ para todo o $l \neq l^*$.

2.2.2 Distribuição processo de Dirichlet

Considere-se, agora, o caso em que o espaço amostral não é finito (por exemplo, a recta real $\mathbb{R}$, o intervalo de 0 a 1, ou um espaço real k-dimensional $\mathbb{R}^k$). Na análise bayesiana paramétrica tradicional, a distribuição de $X_i(G)$ pertenceria a uma família particular de distribuições paramétricas. Por exemplo, se $X_i \in \mathbb{R}$, a distribuição normal, $N(\mu, \sigma)$, é uma das possíveis famílias de distribuições. A análise bayesiana prosseguia considerando os parâmetros, μ e σ, desconhecidos e, consequentemente, atribuindo-lhes uma distribuição *a priori* e, seguidamente, obtendo a correspondente distribuição *a posteriori* para os parâmetros. No entanto, na abordagem bayesiana não paramétrica a distribuição G é considerada desconhecida e, como tal, pretende-se atribuir-lhe uma distribuição *a priori*.

A extensão da distribuição Dirichlet para espaços infinitos é conhecida por distribuição processo de Dirichlet (ou simplesmente processo de Dirichlet), e a definição do processo de Dirichlet é simplificada atendendo à última propriedade apresentada da distribuição Dirichlet.

O processo de Dirichlet (DP) foi introduzido por Ferguson (1973) como uma medida de probabilidade aleatória sobre o espaço de medidas de probabilidades definidas sobre um espaço mensurável $(\Omega, \mathcal{B})$, onde Ω é um espaço amostral não numerável e $\mathcal{B}$ é uma σ-álgebra de subconjuntos de Ω. Um processo de Dirichlet é definido por um parâmetro de concentração, $c > 0$, e uma distribuição de referência sobre o espaço amostral mensurável, G_0. Tal processo pode ser visto como uma distribuição sobre medidas de probabilidade, isto é, cada realização de um processo de Dirichlet é, por si só, uma distribuição de probabilidade.

Diz-se que G é distribuída segundo um processo de Dirichlet se, para qualquer partição mensurável $(B_1, B_2, \ldots, B_k)$ do espaço amostral, o vector de probabilidades $(G(B_1), G(B_2), \ldots, G(B_k))$ é distribuído segundo uma distribuição Dirichlet com vector de parâmetros $(cG_0(B_1), cG_0(B_2), \ldots, cG_0(B_k))$. A notação $G|c, G_0 \sim \mathrm{DP}(c, G_0)$ será

utilizada para indicar que G é definida por um processo de Dirichlet com parâmetros c e G_0.

A distribuição de referência, G_0, pode ser interpretada como a distribuição média do processo uma vez que $G(\cdot) \sim \text{Be}(cG_0(\cdot), c(1 - G_0(\cdot)))$ e, portanto,

$$E[G(\cdot)] = G_0(\cdot).$$

O parâmetro c é referido como o parâmetro de precisão porque controla a variância do processo pois

$$Var[G(\cdot)] = \frac{G_0(\cdot)[1 - G_0(\cdot)]}{c + 1},$$

ou seja, a interpretação dos parâmetros c e G_0 apresentada na secção 2.2.1 continua válida para a presente situação.

Para uma amostra aleatória $X_1, X_2, \ldots, X_n$ escreve-se

$$X_1, X_2, \ldots, X_n | G \overset{\text{iid}}{\sim} G$$
$$G | c, G_0 \sim \text{DP}(c, G_0).$$

A propriedade de conjugação ainda é válida para o processo de Dirichlet, isto é, se $x = (x_1, x_2, \ldots, x_n)$ for a concretização de n variáveis aleatórias independentes e identicamente distribuídas a G, a distribuição *a posteriori* de G é ainda um processo de Dirichlet e pode representar-se por

$$G | x \sim \text{DP}(c + n, G_0^*),$$

onde $G_0^* = \dfrac{cG_0 + nG_n}{c + n}$ e G_n é a função de distribuição empírica das observações.

Vários autores apresentam formas alternativas de definir o processo de Dirichlet: Blackwell e MacQueen (1973) apresentam o processo de Dirichlet como uma representação em esquema de urna de Pólya e Rolin (1992) e Sethuraman (1994) apresentam a denominada definição construtiva do processo de Dirichlet, que é uma representação como uma soma ponderada de massas pontuais. Cada uma das representações fornece um método para gerar realizações de um processo de Dirichlet e ambas são importantes

na implementação dos algoritmos MCMC e na utilização do método de amostragem *Gibbs*.

O processo de Dirichlet requer a especificação de uma distribuição de referência, G_0. Antoniak (1974) sugere centrar o processo de Dirichlet numa família de distribuições paramétricas, com o objectivo de incorporar a família paramétrica na ampla classe de modelos para G. Surge, então, o modelo de mistura de processos de Dirichlet (*Mixture of Dirichlet Processes*) que é especificado simbolicamente como:

$$\begin{aligned} X_1, X_2, \ldots, X_n | G &\overset{\text{iid}}{\sim} G \\ G | c, G_\theta &\sim \text{DP}(c, G_\theta) \quad , \\ \theta &\sim h(\theta) \end{aligned}$$

onde $\{G_\theta : \theta \in \Theta\}$ é uma família paramétrica de modelos probabilísticos.

A simplicidade das suas propriedades e a facilidade de amostrar de um processo de Dirichlet, fizeram com que o modelo se tornasse atractivo e alvo de uma forte investigação nestas duas últimas décadas. No entanto, o processo de Dirichlet ou uma mistura de processos de Dirichlet gera quase certamente distribuições discretas (Ferguson, 1973), limitando a sua aplicação a muitos dos problemas estatísticos, nomeadamente nos problemas de modelação de dados contínuos. Escobar (1994) e Escobar e West (1995) propõem a utilização do processo de Dirichlet noutro contexto, como seja na estimação de densidades. Em particular, os referidos autores utilizam misturas de distribuições normais para estimar densidades e consideram a distribuição da mistura uma quantidade aleatória. A incerteza sobre a distribuição da mistura é descrita utilizando um processo de Dirichlet. Isto é, utilizam o processo de Dirichlet como distribuição *a priori* não paramétrica para a distribuição da mistura.

Considere-se que as variáveis aleatórias X_i, para $i = 1, 2, \ldots, n$, são independentes e provêm de uma mistura de distribuições contínuas, dado o valor de um parâmetro específico, $X_i \sim f(x_i|\theta_i)$, $\theta_i \in \Theta$, onde Θ define um espaço paramétrico. Seja G uma medida de probabilidade aleatória sobre Θ e suponha-se que $G|c, G_0 \sim \text{DP}(c, G_0)$. Então, o modelo denominado modelo de mistura por processo de Dirichlet (*Dirichlet*

Process Mixture) que é especificado simbolicamente como:

$$\begin{aligned} X_i|\theta_i &\overset{\text{ind}}{\sim} f(x_i|\theta_i), \text{ para } i = 1, 2, \ldots, n \\ (\theta_1, \theta_2, \ldots, \theta_n)|G &\overset{\text{iid}}{\sim} G \\ G|c, G_0 &\sim \text{DP}(c, G_0) \end{aligned} \quad ,$$

pode gerar distribuições contínuas e, portanto, colmatar o problema do processo de Dirichlet. O modelo de mistura por processo de Dirichlet é especialmente útil na modelação de dados agrupados em *clusters*, mas é muitas vezes substituído por um processo de Dirichlet, mais por razões matemáticas e computacionais do que por considerações práticas (Hanson, 2006). Uma alternativa não paramétrica bastante flexível aos modelos de mistura por processo de Dirichlet são as distribuições árvores de Pólya e as distribuições mistura de árvores de Pólya, que se apresentam na secção seguinte. Das referências existentes sobre estas últimas distribuições, destaca-se Hanson e Johnson (2002), Paddock et al. (2003) e Hanson (2006).

2.2.3 Distribuições árvores de Pólya

As distribuições árvores de Pólya (PT) definem uma outra distribuição *a priori* não paramétrica, isto é, formam uma classe de distribuições para a medida de probabilidade aleatória G. Estas distribuições são uma generalização do processo de Dirichlet. Em particular, permitem a modelação de distribuições contínuas ou absolutamente contínuas, contornando o problema da discretização do processo de Dirichlet. Foi inicialmente discutida por Freedman (1963), Fabius (1964) e Ferguson (1974) como uma distribuição *tail free*. No entanto, a sua aplicação prática só foi possível mais tarde, nomeadamente depois da introdução dos métodos MCMC, tal como para o processo de Dirichlet. Lavine (1992, 1994) e Mauldin et al. (1992) desenvolveram e catalogaram detalhadamente a base teórica das distribuições árvores de Pólya, apresentando várias aplicações práticas.

Uma distribuição árvore de Pólya para G é construída dividindo o espaço amostral

Ω em intervalos disjuntos cada vez mais pequenos, utilizando o particionamento binário em árvore e atribuindo probabilidades aleatórias a cada um dos ramos da árvore. Teoricamente, o particionamento binário em árvore pode ter infinitos níveis; vai-se restringir esta apresentação às distribuições árvores de Pólya parcialmente especificadas, isto é, finitas e com M níveis. Lavine (1994) discute duas formas de especificar os M níveis. Hanson e Johnson (2002) sugerem a "regra de ouro" $M \approx \log_2(n)$, permitindo que o número de níveis M aumente com o aumento da dimensão da amostra.

Seja $\{B_0, B_1\}$ uma partição mensurável de Ω, no primeiro nível. Segue-se no segundo nível $\{B_{00}, B_{01}\}$ uma partição mensurável de B_0 e $\{B_{10}, B_{11}\}$ uma partição mensurável de B_1. Continua-se o particionamento binário da árvore até atingir M níveis (i.e. $m = 1, 2, \ldots, M$), sendo o conjunto de todas as partições binárias, uma sequência finita de partições binárias em árvore de Ω. Considere-se, no m-ésimo nível, $\varepsilon_{1:m} = \varepsilon_1\varepsilon_2\cdots\varepsilon_m$ com cada $\varepsilon_j \in \{0, 1\}$, para $j = 1, 2, \ldots, m$, tal que cada $\varepsilon_{1:m}$ define uma única partição $B_{\varepsilon_{1:m}}$. O número de partições binárias no m-ésimo nível é 2^m e tem-se $B_{\varepsilon_{1:m}}$ dividido em $B_{\varepsilon_{1:m}0}$ e $B_{\varepsilon_{1:m}1}$ no nível $(m+1)$, tal que $\Omega = B_0 \cup B_1$, $B_0 \cap B_1 = \emptyset$ e para cada $\varepsilon_{1:m}$, $B_{\varepsilon_{1:m}} = B_{\varepsilon_{1:m}0} \cup B_{\varepsilon_{1:m}1}$ e $B_{\varepsilon_{1:m}0} \cap B_{\varepsilon_{1:m}1} = \emptyset$. Denomine-se por $\Pi = \{B_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$ uma sequência finita de partições binárias em árvore de Ω.

Para definir uma medida de probabilidade aleatória sobre Ω atribui-se medidas de probabilidade aleatórias à sequência finita de partições binárias $B_{\varepsilon_{1:m}}$, para $m = 1, 2, \ldots, M$. Partindo de Ω, uma observação pertence a B_0 com probabilidade Y_0, ou pertence a B_1 com probabilidade $Y_1 = 1 - Y_0$. No nível 2, por exemplo, uma observação pertence a B_{00} dado que pertence a B_0 com probabilidade Y_{00}. Generalizando para $m \geq 2$, ao entrar em $B_{\varepsilon_{1:m}}$ uma observação pode mover-se para $B_{\varepsilon_{1:m}0}$ com uma probabilidade condicional $Y_{\varepsilon_{1:m}0}$ ou mover-se para $B_{\varepsilon_{1:m}1}$ com uma probabilidade condicional $Y_{\varepsilon_{1:m}1} = 1 - Y_{\varepsilon_{1:m}0}$. A distribuição marginal de uma sequência $B_{\varepsilon_{1:m}}$, no m-ésimo nível, é dada por

$$G(B_{\varepsilon_{1:m}}) = \left(\prod_{j=1,\varepsilon_j=0}^{m} Y_{\varepsilon_1\cdots\varepsilon_{j-1}0}\right)\left(\prod_{j=1,\varepsilon_j=1}^{m} (1 - Y_{\varepsilon_1\cdots\varepsilon_{j-1}0})\right),$$

onde, para o primeiro nível, i.e. para $j = 1$, se tem Y_0 ou $1 - Y_0$.

Por exemplo, para $m = 2$, $G(B_{00}) = Y_0 Y_{00}$, $G(B_{01}) = Y_0(1 - Y_{00})$, $G(B_{10}) = (1 - Y_0)Y_{10}$ e $G(B_{11}) = (1 - Y_0)(1 - Y_{10})$. Por definição, $Y_{(\cdot)}$ são variáveis aleatórias independentes com distribuição beta, isto é, $Y_0 \sim \text{Be}(\alpha_0, \alpha_1)$ e para todo o $\varepsilon_{1:m}$, $Y_{\varepsilon_{1:m}0} \overset{\text{ind}}{\sim} \text{Be}(\alpha_{\varepsilon_{1:m}0}, \alpha_{\varepsilon_{1:m}1})$, com parâmetros α_0, α_1, $\alpha_{\varepsilon_{1:m}0}$ e $\alpha_{\varepsilon_{1:m}1}$ não negativos. Denomine-se por $\mathcal{A} = \{\alpha_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$ o conjunto de parâmetros não negativos.

Uma distribuição árvore de Pólya finita com M níveis é determinada pelas partições em Π e pelos parâmetros da distribuição beta em $\mathcal{A}$ e representa-se por $G|\Pi, \mathcal{A} \sim \text{PT}_M(\Pi, \mathcal{A})$.

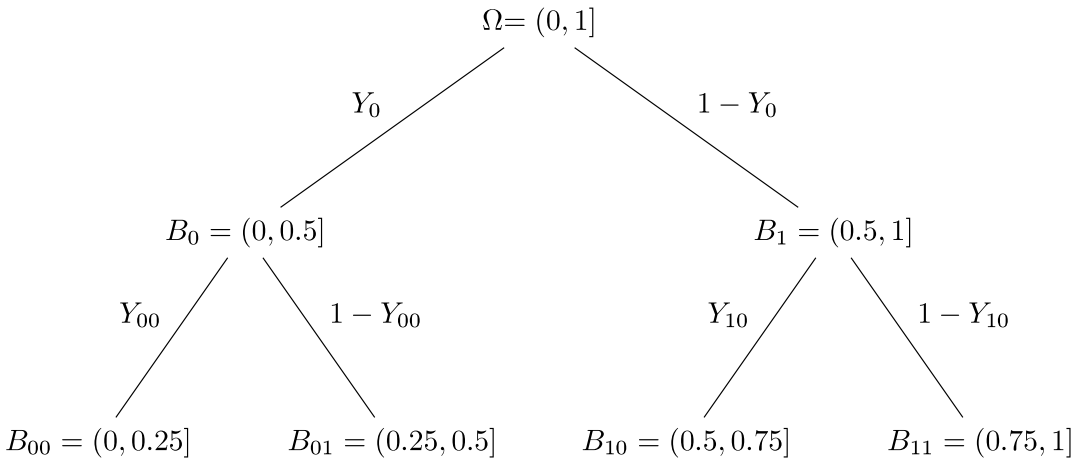

Figura 2.2: Ilustração de uma distribuição árvore de Pólya com dois níveis para uma partição binária do espaço amostral $\Omega = (0, 1]$.

Para ilustração, apresenta-se, na Figura 2.2, a construção de uma árvore de Pólya finita com dois níveis. O espaço amostral Ω (que se fixou no intervalo $(0, 1]$ como em Ferguson (1974)) é dividido numa sequência encaixada de partições binárias com $M = 2$ níveis. No primeiro nível, Ω é dividido em dois intervalos, $B_0 = (0, 0.5]$ e $B_1 = (0.5, 1]$, tal que $\Omega = B_0 \cup B_1$ com $B_0 \cap B_1 = \emptyset$ e Y_0 e $1 - Y_0$ representam a probabilidade de uma observação pertencer a B_0 e a B_1, respectivamente. Seguem-se as partições encaixadas no segundo nível, $B_{00} = (0, 0.25]$ e $B_{01} = (0.25, 0.5]$ uma partição de B_0, e $B_{10} = (0.5, 0.75]$ e $B_{11} = (0.75, 1]$ uma partição de B_1, onde, agora, Y_{00}, $1 - Y_{00}$, Y_{10} e $1 - Y_{10}$ são, respectivamente, as probabilidades de uma observação pertencer a cada

um dos intervalos, dado que pertence ao correspondente intervalo que lhe deu origem.

Se as variáveis aleatórias $X_1, X_2, \ldots, X_n$ constituem uma amostra de dimensão n de G, e G é definida por uma distribuição árvore de Pólya finita com M níveis, então, simbolicamente, tem-se que

$$X_1, X_2, \ldots, X_n | G \overset{\text{iid}}{\sim} G$$
$$G|\Pi, \mathcal{A} \sim \text{PT}_M(\Pi, \mathcal{A}).$$

A propriedade de conjugação também se verifica neste caso (Lavine, 1992) e é dada por

$$G|x \sim \text{PT}_M(\Pi, \mathcal{A}^*),$$

onde $\mathcal{A}^* = \{\alpha^*_{\varepsilon_{1:m}} = \alpha_{\varepsilon_{1:m}} + n_{\varepsilon_{1:m}}\}$ e $n_{\varepsilon_{1:m}}$ é o número de observações na amostra que pertencem a $B_{\varepsilon_{1:m}}$.

Na prática, para determinar as partições, $\Pi = \{B_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$, e eliciar os parâmetros da distribuição beta, $\mathcal{A} = \{\alpha_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$, Lavine (1992) propõe uma construção canónica da distribuição árvore de Pólya sobre $\Omega \subset \mathbb{R}$, com função distribuição $G_0(\cdot)$. Esta construção centra a distribuição árvore de Pólya em torno de uma distribuição G_0 e considera que os limites (inferior e superior) dos conjuntos que constituem cada nível m da partição, coincidem com os quantis de G_0 tal que $\alpha_{\varepsilon_{1:m}0} = \alpha_{\varepsilon_{1:m}1}$, para todo o $\varepsilon_{1:m}$. Por exemplo, se $X_i \in \mathbb{R}$ tem-se que no primeiro nível $B_0 = (-\infty, G_0^{-1}(0.5)]$ e $B_1 = (G_0^{-1}(0.5), -\infty)$. Generalizando, no m-ésimo nível, os conjuntos que formam a partição são definidos através dos seguintes intervalos

$$B_{\varepsilon_{1:m}} = \left\{ \left(G_0^{-1}\left(\frac{k-1}{2^m}\right), G_0^{-1}\left(\frac{k}{2^m}\right) \right] \right\},$$

para $m = 1, 2, \ldots, M$ e $k = 1, 2, \ldots, 2^m$, onde $G_0^{-1}(0) = -\infty$ e $G_0^{-1}(1) = +\infty$.

Desta forma, por exemplo, como $G(B_0) = Y_0 \sim \text{Be}(\alpha_0, \alpha_1)$ e considerando que $\alpha_0 = \alpha_1$, tem-se que

$$E[G(B_0)] = E[Y_0] = \frac{\alpha_0}{\alpha_0 + \alpha_1} = 1/2 = G_0(B_0).$$

Generalizando, tem-se que $E[G(B_{\varepsilon_{1:m}})] = 2^{-m} = G_0(B_{\varepsilon_{1:m}})$. Então G_0 tem um comportamento semelhante ao da distribuição base de um processo de Dirichlet.

Ao centrar a distribuição árvore de Pólya em torno da medida de probabilidade G_0, o conjunto de parâmetros da distribuição beta, $\mathcal{A} = \{\alpha_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$, determina o quanto G está concentrada perto da sua média, isto é G_0. Além disso, os elementos do conjunto $\mathcal{A}$ podem ser utilizados para representar as crenças *a priori*, sendo usual definilos através de constantes apropriadas, seja $\alpha_{\varepsilon_{1:m}} = c_m$, para todo o $\varepsilon_{1:m}$ no nível m. Estas constantes vão ter como função controlar a continuidade de G e as condições que fazem com que G seja contínua requerem que os c_m aumentem rapidamente, isto é, que a variância seja reduzida rapidamente à medida que se vai descendo na árvore. Segundo Ferguson (1974), $c_m = m^2$ implica que G é absolutamente contínua com probabilidade um e, portanto, de acordo com Lavine (1992), esta será uma "escolha canónica sensata". Walker e Mallick (1997) e Paddock (1999) consideram $c_m = cm^2$, com $c > 0$. Como m é fixo, então c é um parâmetro de concentração, isto é, para valores grandes de c a distribuição *a priori* está mais concentrada em torno de G_0. Por exemplo, as medidas de probabilidade aleatórias, G, geradas de uma distribuição árvore de Pólya estarão mais concentradas, tanto em termos de semelhança na forma como em termos de distância, em torno da medida de probabilidade G_0. Se c está próximo de zero, as medidas de probabilidade aleatórias geradas estarão consideravelmente mais dispersas, quer em termos de forma quer em termos de distância, relativamente à medida de probabilidade G_0. Se $c_m = c/2^m$, cai-se num processo de Dirichlet, o que significa que $c_m \to 0$ à medida que $m \to \infty$ e G é discreta com probabilidade um (Blackwell e MacQueen, 1973). Berger e Guglielmi (2001) consideram $c_m = c\rho(m)$, para $\rho(m) = m^2$, m^3, 2^m, 4^m e 8^m. Em particular, qualquer $\rho(m)$ tal que $\sum_{m=1}^{\infty} \rho(m)^{-1} < \infty$, garante que G seja absolutamente contínua (Schervish, 1995).

Conclui-se, portanto, que a distribuição árvore de Pólya é mais geral e inclui o processo de Dirichlet como caso particular, tendo como principal vantagem gerar medidas de probabilidade aleatórias para variáveis contínuas. No entanto, as distribuições ár-

vores de Pólya, definidas até aqui, têm algumas limitações práticas, tais como: (i) a medida de probabilidade aleatória é dependente da partição que for considerada; (ii) Barron et al. (1999) observaram que a densidade preditiva *a posteriori* apresenta, em geral, descontinuidades nos pontos extremos dos intervalos que definem as partições; e (iii) a inerente dificuldade na escolha da medida de probabilidade G_0.

Para contornar estas dificuldades, e seguindo a sugestão de Lavine (1992), Hanson e Johnson (2002) e Hanson (2006) propõem que a medida de probabilidade G_0 possa depender de parâmetros desconhecidos (hiperparâmetros), ou seja, possa ser definida por uma medida de probabilidade paramétrica, por exemplo G_θ, e considere-se distribuições *a priori*, $h(\theta)$, para esses hiperparâmetros. Desta forma, tem-se uma família de medidas de probabilidade $\{G_\theta : \theta \in \Theta\}$ e, consequentemente, uma família de partições $\{\Pi_\theta : \theta \in \Theta\}$. Segundo estes autores, este procedimento suaviza as descontinuidades nos pontos extremos dos intervalos, originando, assim, densidades preditivas *a posteriori* mais suaves. O modelo resultante é designado por mistura finita de árvores de Pólya (MPT) com M níveis e é frequentemente representado pela seguinte estrutura hierárquica

$$\begin{aligned} X_1, X_2, \ldots, X_n | G &\overset{\text{iid}}{\sim} G \\ G | \Pi_\theta, \mathcal{A} &\sim \text{MPT}_M(\Pi_\theta, \mathcal{A}) \\ \theta &\sim h(\theta). \end{aligned}$$

Resumindo, a construção apresentada para definir o modelo mistura finita de árvores de Pólya considera que apenas as partições da árvore dependem de θ, Π_θ, e a família de parâmetros $\mathcal{A} = \{\alpha_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$ são fixos, por exemplo, $\alpha_{\varepsilon_{1:m}} = m^2$. Outra construção possível para um modelo mistura finita de árvores de Pólya consiste em manter as partições fixas e fazer os parâmetros da família $\mathcal{A}$ depender de θ, tal que se mantenha a relação $E[G(B_{\varepsilon_{1:m}})|\theta] = G_\theta(B_{\varepsilon_{1:m}})$. Esta última construção foi adoptada por Berger e Guglielmi (2001) para testar uma família paramétrica contra uma família alternativa não paramétrica. Mais detalhes sobre esta construção serão apresentados mais à frente aquando da introdução do teste para o estudo da adequabilidade de um

modelo paramétrico. A representação deste modelo é

$$X_1, X_2, \ldots, X_n | G \overset{\text{iid}}{\sim} G$$
$$G | \Pi, \mathcal{A}_\theta \sim \text{MPT}_M(\Pi, \mathcal{A}_\theta)$$
$$\theta \sim h(\theta).$$

2.3 Critérios de comparação de modelos

2.3.1 Factor de Bayes

Na abordagem bayesiana, o factor de Bayes (BF, *Bayes Factor*), introduzido por Jeffreys (1961), é um dos critérios de eleição para a comparação de modelos. O teste pode ser escrito na forma

$$H_0 : X \sim M_1 \qquad vs \qquad H_1 : X \sim M_2,$$

onde se supõe, por simplicidade na exposição, que ambos os modelos são paramétricos.

O factor de Bayes é uma medida da evidência provida pelos dados a favor de uma das hipóteses (modelos) em confronto.

Sejam $f_j(x|\theta_j)$ e $h_j(\theta_j)$, a distribuição amostral dado o vector de parâmetros e a distribuição *a priori* para o vector de parâmetros, sob o modelo M_j, para $j = 1, 2$, respectivamente. Represente-se por $\Pr\{M_j\}$ a probabilidade *a priori* de M_j ser o modelo verdadeiro, com $\Pr\{M_2\} = 1 - \Pr\{M_1\}$.

Uma vez observado o conjunto de dados x, o teorema de Bayes é utilizado para obter a probabilidade *a posteriori* de M_j ser o modelo verdadeiro,

$$\Pr\{M_j|x\} = \frac{p_j(x)\Pr\{M_j\}}{p_1(x)\Pr\{M_1\} + p_2(x)\Pr\{M_2\}}, \text{ para } j = 1, 2,$$

onde $p_j(x) = \int f_j(x|\theta_j) h_j(\theta_j) d\theta_j$ define a distribuição preditiva *a priori* ou marginal de X, para o modelo M_j.

O factor de Bayes a favor de M_2 e contra M_1 é definido como o quociente entre a razão das vantagens *a posteriori* e a razão das vantagens *a priori* e representa-se por

$$\mathrm{BF}^{21}(x) = \frac{\Pr\{M_2|x\}}{\Pr\{M_1|x\}} / \frac{\Pr\{M_2\}}{\Pr\{M_1\}},$$

que se pode escrever simplesmente como

$$\mathrm{BF}^{21}(x) = \frac{p_2(x)}{p_1(x)} = \frac{\int f_2(x|\theta_2)h_2(\theta_2)d\theta_2}{\int f_1(x|\theta_1)h_1(\theta_1)d\theta_1}. \tag{2.2}$$

Intuitivamente, o melhor modelo corresponde àquele que apresente o maior valor da distribuição preditiva *a priori* para x. Um factor de Bayes muito grande ou muito pequeno relativamente a um representa uma evidência muito forte nos dados a favor de uma hipótese contra a outra hipótese. Assim, se $\mathrm{BF}^{21}(x) > 1$ os dados x favorecem M_2, e se $\mathrm{BF}^{21}(x) < 1$ os dados x favorecem M_1. No entanto, é usual determinar um valor de corte (*threshold*) para o factor de Bayes que permita tomar uma decisão a favor de um modelo. Normalmente, utilizam-se os valores recomendados por Jeffreys (1961), que são apresentados na Tabela 2.1.

O factor de Bayes definido em (2.2), que se vai também designar de factor de Bayes simples quando necessário, pode ser interpretado como a razão das médias das verosimilhanças dos parâmetros para os dois modelos, calculadas relativamente às respectivas distribuições *a priori* do vector de parâmetros, ou seja, a razão das médias *a priori* das verosimilhanças. Pode assim dizer-se que este factor de Bayes tem alguma semelhança com o teste da razão de verosimilhanças clássico, com a diferença de que no cálculo do factor de Bayes o vector de parâmetros é eliminado por integração enquanto que no teste da razão de verosimilhanças o vector de parâmetros é substituído pelos estimadores de máxima verosimilhança.

Jeffreys (1961) sugeriu interpretar o factor de Bayes dividindo os possíveis valores do seu logaritmo de base 10 ($\log_{10}(.)$, por forma a ter valores mais pequenos) em vários intervalos que são apresentados na Tabela 2.1; nesta tabela é, ainda, apresentada a interpretação para cada intervalo de valores.

Kass e Raftery (1995) propõem fazer a divisão de acordo com a Tabela 2.2, isto é, considerando para valores de corte do factor de Bayes duas vezes o seu logaritmo

Tabela 2.1: Interpretação dos valores do factor de Bayes (Jeffreys, 1961).

$\mathrm{BF}^{21}(x)$	$\log(\mathrm{BF}^{21}(x))$	Evidência a favor de M_2
< 1	< 0	negativa (favorece M_1)
1 a 3.2	0 a 0.5	insignificante
3.2 a 10	0.5 a 1	significativa
10 a 100	1 a 2	forte
> 100	> 2	muito forte

neperiano. Tem-se, desta forma, que o valor obtido fica na mesma escala do teste de razão de verosimilhanças.

Tabela 2.2: Interpretação dos valores do factor de Bayes (Kass e Raftery, 1996).

$\mathrm{BF}^{21}(x)$	$2\ln(\mathrm{BF}^{21}(x))$	Evidência a favor de M_2
< 1	< 0	negativa (favorece M_1)
1 a 3	0 a 2	insignificante
3 a 20	2 a 6	significativa
20 a 150	6 a 10	forte
> 150	> 10	muito forte

Alguns autores optam por definir o factor de Bayes a favor de M_0 ($H_0 : X \sim M_0$) e contra M_1 ($H_1 : X \sim M_1$). Neste caso, o factor de Bayes é representado por $\mathrm{BF}_{01}(x)$.

Uma das limitações do factor de Bayes simples reside no facto de este depender das distribuições *a priori* do vector de parâmetros dos dois modelos em comparação, $h_j(\theta_j)$, para $j = 1, 2$. Usualmente, são utilizadas distribuições *a priori* não informativas, podendo originar distribuições *a priori* impróprias e, consequentemente, as correspondentes distribuições preditivas *a priori* também podem ser impróprias e, tem-se, deste

modo, inviabilizado o próprio cálculo do factor de Bayes simples. De modo a contornar o problema da especificação de distribuições que representem situações de ignorância *a priori*, vários autores sugerem modificações ao factor de Bayes simples.

A segunda limitação tem a ver com questões computacionais. Como o factor de Bayes é função das densidades preditivas *a priori*, $p_j(x)$, para $j = 1, 2$, o seu cálculo analítico, em geral, só é possível em situações ou modelos simples, como é o caso das distribuições da família exponencial com distribuições *a priori* conjugadas (DeGroot, 1970).

Quando as distribuições preditivas *a priori* existem mas os integrais que as definem são difíceis de resolver analiticamente, são utilizados métodos para a aproximação do factor de Bayes simples, nomeadamente métodos baseados em aproximações analíticas e aproximações numéricas como, por exemplo, o método de Laplace de aproximação de integrais e a quadratura iterativa, respectivamente. Os métodos de Monte Carlo também são uma alternativa apropriada aos métodos numéricos para aproximação de integrais, nomeadamente o método de Monte Carlo com amostragem via função de importância. Todos estes métodos podem ser vistos com grande detalhe em Paulino et al. (2003), capítulos 5 e 7.

As modificações ao factor de Bayes simples, que se apresentam seguidamente, permitem o uso de distribuições *a priori* não informativas, por vezes impróprias, e também podem ser interpretadas usando as regras apresentadas nas Tabelas 2.1 e 2.2.

Factores de Bayes alternativos

Aitkin (1991) propõe a substituição das distribuições *a priori*, $h_j(\theta_j)$, no factor de Bayes simples definido em (2.2), pelas correspondentes distribuições *a posteriori*, $h_j(\theta_j|x)$, para $j = 1, 2$. Surge deste modo o factor de Bayes *a posteriori*, que é dado por

$$\mathrm{BF}^{21}_{post}(x) = \frac{p_2^{post}(x)}{p_1^{post}(x)} = \frac{\int f_2(x|\theta_2)h_2(\theta_2|x)d\theta_2}{\int f_1(x|\theta_1)h_1(\theta_1|x)d\theta_1},$$

ou seja, é a razão das médias *a posteriori* das verosimilhanças para os dois modelos. Note-se que é apenas necessário garantir que $h_j(\theta_j|x)$, para $j = 1, 2$, seja própria quando se utilizam distribuições *a priori* impróprias.

O'Hagan (1991, 1995) critica o factor de Bayes *a posteriori*, devido ao facto dos dados serem utilizados duas vezes, primeiro na determinação da distribuição *a posteriori* e depois no cálculo do factor de Bayes, propondo a sua substituição pelo factor de Bayes parcial, que se apresenta a seguir.

A ideia do factor de Bayes parcial consiste em dividir o conjunto de dados, isto é, a amostra completa, em duas partes, $x = (x^{(1)}, x^{(2)})$. Uma parte do conjunto de dados, designada por amostra de treino, $x^{(1)}$, é utilizada para actualizar a distribuição *a priori*, ou seja, para obter a distribuição *a posteriori*, e a outra parte do conjunto de dados (as observações restantes), $x^{(2)}$, é utilizada para calcular o factor de Bayes. O factor de Bayes parcial é, assim, dado por

$$\mathrm{BF}^{21}_{parc}(x^{(2)}|x^{(1)}) = \frac{p_2(x^{(2)}|x^{(1)})}{p_1(x^{(2)}|x^{(1)})} = \frac{\int f_2(x^{(2)}|\theta_2)h_2(\theta_2|x^{(1)})d\theta_2}{\int f_1(x^{(2)}|\theta_1)h_1(\theta_1|x^{(1)})d\theta_1}.$$

O factor de Bayes parcial, é referido por O'Hagan (1995) como sendo menos sensível à escolha das distribuições *a priori*, ultrapassa o problema das distribuições *a priori* impróprias, mas apresenta um novo problema: como dividir o conjunto de dados x em duas partes?

Uma solução possível é dada por Berger e Pericchi (1993, 1996). Estes autores sugerem utilizar todas as amostras de treino de dimensão mínima para actualizar as distribuições *a priori* e determinar a média aritmética (ou geométrica) dos factores de Bayes parciais. Isto é, utilizar todos os sub-conjuntos de dados de dimensão n_1 (amostras de x), onde n_1 representa a menor dimensão da amostra que conduz a distribuições *a posteriori* próprias, calculando posteriormente a média dos correspondentes factores de Bayes. Esta média é denominada de factor de Bayes intrínseco (aritmético ou geométrico).

Para O'Hagan (1995, 1997), a redução na sensibilidade à escolha da distribuição *a*

priori, do factor de Bayes intrínseco, é mínima quando são utilizadas distribuições *a priori* impróprias, e a redução é nula quando são utilizadas distribuições *a priori* próprias. O mesmo autor propõe, evitando a escolha arbitrária de uma amostra de treino ou ter que considerar todos os sub-conjuntos de dados de dimensão n_1, utilizar uma proporção do conjunto de dados x, definida por $b = n_1/n$, onde n_1 é a já mencionada dimensão de amostra mínima e n é a dimensão da amostra completa. Suponde n_1 e n grandes O'Hagan (1995) conclui que a verosimilhança $f_j(x^{(1)}|\theta_j)$, baseada apenas na amostra de treino $x^{(1)}$, será aproximadamente igual a $f_j(x|\theta_j)^b$. Surge assim o factor de Bayes fraccionário dado por

$$\mathrm{BF}^{21}_{frac}(x;b) = \frac{p_2(x;b)}{p_1(x;b)}, \tag{2.3}$$

onde

$$p_j(x;b) = \frac{\int f_j(x|\theta_j)h_j(\theta_j)d\theta_j}{\int f_j(x|\theta_j)^b h_j(\theta_j)d\theta_j}, \text{ para } j = 1,2. \tag{2.4}$$

Suponha-se que $h_j(\theta_j)$, para $j = 1,2$, são distribuições *a priori* impróprias, ou seja,

$$h_j(\theta_j) \propto g_j(\theta_j), \text{onde} \int g_j(\theta_j)d\theta_j \to \infty, \text{ para } j = 1,2.$$

Formalmente, pode escrever-se $h_j(\theta_j) = c_j^{te} g_j(\theta_j)$, embora as constantes de normalização c_j^{te} não existam, mas tratando-as como constantes arbitrárias. Consequentemente, a expressão definida em (2.4) é substituída por

$$p_j(x;b) = \frac{\int f_j(x|\theta_j)c_j^{te}g_j(\theta_j)d\theta_j}{\int f_j(x|\theta_j)^b c_j^{te} g_j(\theta_j)d\theta_j}, \text{ para } j = 1,2,$$

e como as constantes arbitrárias c_j^{te} não dependem de θ_j, respectivamente, estas cancelam-se. Desta forma, e desde que os integrais envolvidos sejam convergentes, evita-se que o factor de Bayes fraccionário seja indefinido.

Outros autores (Geisser e Eddy, 1979 e Gelfand et al., 1992) adoptam uma metodologia de validação cruzada *leave one out* para definir o factor de Bayes, isto é, a amostra completa é dividida em duas partes, $x = (x_{(-i)}, x_i)$, mas, agora, a amostra de treino, $x_{(-i)}$, é constituída por todas as observações à excepção de x_i. Se X_i, para

$i = 1, 2, \ldots, n$, forem condicionalmente independentes dado θ, a distribuição preditiva *a priori*, $p_j(x)$, é substituída pela denominada distribuição pseudo-preditiva, dada por

$$\prod_{i=1}^{n} p_j(x_i|x_{(-i)}),$$

onde,

$$p_j(x_i|x_{(-i)}) = \int_{\Theta} f_j(x_i|\theta_j) h_j(\theta_j|x_{(-i)}) d\theta_j.$$

O valor de $p_j(x_i|x_{(-i)})$, quando x_i é a i-ésima componente da amostra completa x, denomina-se na literatura por ordenada preditiva condicional da observação x_i para o modelo M_j e representa-se usualmente por $\mathrm{CPO}_j(x_i)$ (CPO, *Conditional Predictive Ordinate*). É uma medida muito utilizada como um método de diagnóstico informal de observações mal ajustadas pelo modelo. Como estes valores são um indicador da verosimilhança de cada observação dadas todas as outras observações, valores baixos de $\mathrm{CPO}_j(x_i)$ devem corresponder a observações mal ajustadas pelo modelo (Paulino et al., 2003). Um diagrama de dispersão das $\mathrm{CPO}_j(x_i)$ *versus* a ordem das observações permite detectar rapidamente possíveis observações discrepantes (*outliers*).

Obtém-se assim o chamado factor pseudo-Bayes, dado por

$$\mathrm{BF}^{21}_{pseudo}(x) = \prod_{i=1}^{n} \frac{p_2(x_i|x_{(-i)})}{p_1(x_i|x_{(-i)})} = \prod_{i=1}^{n} \frac{\mathrm{CPO}_2(x_i)}{\mathrm{CPO}_1(x_i)}. \tag{2.5}$$

A principal motivação para o uso das distribuições preditivas condicionais, $p(x_i|x_{(-i)})$, para $i = 1, 2, \ldots, n$, reside no facto de que esta existe, isto é, é própria, se $h(\theta|x_{(-i)})$ também o for, permitindo assim o uso de distribuições *a priori* não informativas.

O principal problema no uso das distribuições preditivas condicionais está na dificuldade computacional, uma vez que a distribuição preditiva tem de ser calculada para diferentes conjuntos de dados à medida que i varia. Gelfand (1996) sugere, no entanto, a utilização do método MCMC para a estimação dessas distribuições preditivas (ver Paulino et al. (2003), secção 8.4.3).

2.3.2 Critério de informação da *deviance*

Um outro critério de comparação de modelos baseado na verosimilhança, com penalizações impostas ao incremento do número de parâmetros no modelo, é o critério de informação da *deviance* (DIC, *Deviance Information Criterion*) proposto por Spiegelhalter et al. (2002).

Considere-se $x = (x_1, x_2, \ldots, x_n)$ uma amostra observada que pode ser associada a um de r modelos de probabilidade, M_j, para $j = 1, 2, \ldots, r$. A cada um dos modelos tem-se, respectivamente, associada uma função densidade ou função de probabilidade com vector de parâmetros θ_j, $f_j(x|\theta_j)$, e uma função de verosimillhança $L_j(\theta_j|x) = \prod_{i=1}^{n} f_j(x_i|\theta_j)$.

A *deviance* bayesiana do modelo M_j, é dada por

$$D_j(\theta_j) = -2\ln\left[L_j(\theta_j|x)\right] + 2\ln\left[f_j(x)\right], \tag{2.6}$$

onde $f_j(x)$ é uma função apenas dos dados, que não tem impacto na escolha do modelo. Deste modo, é usual utilizar-se $f_j(x) = 1$ para todos os modelos.

Assim, (2.6) é dada por

$$D_j(\theta_j) = -2\ln\left[L_j(\theta_j|x)\right].$$

Com base neste critério o ajustamento do modelo é sumariado através do valor esperado *a posteriori* da *deviance* bayesiana, $E_{\theta_j|x}[D_j(\theta_j)]$, e a complexidade do modelo é expressa através do número efectivo de parâmetros do modelo, p_{D_j}, que é definido como sendo a diferença entre o valor esperado *a posteriori* da *deviance* bayesiana e o valor da *deviance* bayesiana calculado no valor esperado *a posteriori* de θ_j, isto é

$$p_{D_j} = E_{\theta_j|x}[D_j(\theta_j)] - D_j(E_{\theta_j|x}[\theta_j]). \tag{2.7}$$

A equação (2.7) pode ser reescrita na forma

$$E_{\theta_j|x}[D_j(\theta_j)] = D_j(E_{\theta_j|x}[\theta_j]) + p_{D_j}$$

e, assim, a medida de ajustamento do modelo pode ser interpretada como a soma de uma medida clássica de ajustamento por estimativa directa com uma medida de complexidade do modelo. Por este motivo, o valor esperado *a posteriori* da *deviance*, $E_{\theta_j|x}[D_j(\theta_j)]$, pode ser vista, também, como uma medida de adequabilidade do modelo.

Finalmente, Spiegelhalter et al. (2002) definem o critério de informação da *deviance* que consiste em escolher o modelo M_j que apresente o menor valor de

$$\mathrm{DIC}_j(x) = E_{\theta_j|x}[D_j(\theta_j)] + p_{D_j}. \tag{2.8}$$

O DIC é um critério de fácil implementação via métodos MCMC. Substituindo em (2.8) a expressão de p_{D_j} pela dada em (2.7) e simplificando a notação, tem-se

$$\mathrm{DIC}_j(x) = 2E_{\theta_j|x}[D_j(\theta_j)] - D_j(E_{\theta_j|x}[\theta_j]) = 2\bar{D}_j - D_j(\bar{\theta}_j).$$

Seja $(\theta_j^{(1)}, \theta_j^{(2)}, \ldots, \theta_j^{(L)})$ uma amostra simulada, de dimensão L, da distribuição *a posteriori*, $h_j(\theta_j|x)$; então, tem-se que

$$\bar{D}_j = \frac{1}{L}\sum_{l=1}^{L} D_j(\theta^{(l)})$$

e

$$D_j(\bar{\theta}_j) = D_j\left(\frac{1}{L}\sum_{l=1}^{L}\theta^{(l)}\right).$$

Se o objectivo da utilização do DIC for a comparação de dois modelos, M_1 e M_2, pode usar-se a diferença

$$\Delta\mathrm{DIC}^{12}(x) = \mathrm{DIC}_1(x) - \mathrm{DIC}_2(x),$$

onde $\Delta\mathrm{DIC}^{12}(x)$ representa a alteração do valor do DIC do modelo M_1 para o modelo M_2. Se $\Delta\mathrm{DIC}^{12} < 0$, selecciona-se o modelo M_1, senão selecciona-se o modelo M_2.

Capítulo 3

Métodos de estudo da adequabilidade de modelos

Independentemente da forma de selecção de um determinado modelo, é fundamental averiguar se ele é adequado para realizar as inferências necessárias à obtenção das respostas relevantes de interesse, isto é, é importante validar o modelo. O objectivo não é determinar se o modelo é verdadeiro ou falso, mas sim saber até que ponto as deficiências do modelo interferem no processo de inferência (Gelman et al., 1995). Convém relembrar que o conceito de modelo, a utilizar nos testes de ajustamento globais e específicos, refere-se apenas à distribuição dos dados, condicional a um vector de parâmetros desconhecido, ou seja à distribuição amostral $f(x|\theta)$. Quando se utiliza uma abordagem bayesiana, é também definida uma distribuição *a priori* para o vector de parâmetros.

Várias metodologias bayesianas, para abordar a questão da validação de um determinado modelo, têm sido propostas. Alguns autores sugerem o uso da distribuição preditiva *a priori* para a validação do modelo. Considere-se, para a amostra aleatória $X = (X_1, X_2, \ldots, X_n)$, o modelo paramétrico M: $X|\theta \sim f(x|\theta)$, e $\theta \sim h(\theta)$ com $\theta \in \Theta$.

A distribuição preditiva *a priori* de X, condicional ao modelo M é dada por

$$p(x) = \int_{\Theta} f(x|\theta)h(\theta)d\theta, \tag{3.1}$$

que não é mais do que a constante normalizadora da distribuição *a posteriori*. A distribuição preditiva definida em (3.1) pode ser vista como a esperança *a priori* da função de verosimilhança, dado o modelo M. É muitas vezes designada por verosimilhança preditiva, uma vez que é obtida depois de integrar no correspondente vector de parâmetros.

Segundo Carlin e Louis (2000), valores observados x_i, do vector x, para os quais o valor da distribuição marginal $p(x_i)$ é demasiado pequeno são "improváveis", devendo, portanto, ser considerados valores discrepantes sob a validade do modelo M. Um grande número de valores pequenos de $p(x_i)$ sugere que o modelo M deve ser considerado inadequado e que outros modelos devem ser propostos.

A principal dificuldade desta abordagem é definir a partir de que valor $p(x_i)$ deve ser considerado pequeno. Esta abordagem também tem o inconveniente de não poder ser utilizada quando a distribuição preditiva *a priori* é imprópria, situação que ocorre frequentemente quando se utiliza distribuições *a priori* não informativas. A descrição detalhada deste método, assim como formas de obter estimativas da distribuição preditiva *a priori* $p(x)$, quando ela existe, pode ser encontrada em Paulino et al. (2003, secção 8.4.3).

Para contornar as dificuldades encontradas no uso da distribuição preditiva *a priori*, outras abordagens são sugeridas por diversos autores, nomeadamente métodos que fazem uso da distribuição preditiva *a posteriori* (*e.g.,* Bayarri e Berger, 2000; Gelman et al., 2004).

3.1 Medidas de surpresa

Entende-se por medida de surpresa a quantificação do grau de incompatibilidade dos dados com o modelo proposto, sem recurso a modelos alternativos. Em Bayarri e Berger (1997) é apresentado um estudo exaustivo de diversas medidas de surpresa, sendo dado especial destaque ao uso do valor-p (*p-value*) como medida de surpresa nos dados.

Os mesmos autores referem: "*We do not believe that a "surprise" analysis can ever replace a full Bayesian one. We do argue, however, that surprise measures have an important role to play as exploratory tools, in the sense that, if* x_{obs} *can be nicely explained by* H_0 *we might not need to take extra effort of the full analysis. If, however,* x_{obs} *is "surprising" then we do have to carefully specify alternative models to* H_0 *and carry out a Bayesian analysis*".

Na literatura estatística são apresentadas várias abordagens, quer clássicas quer bayesianas, ao estudo do valor-p (*e.g.,* Bayarri e Berger, 1997, 2000; Bayarri e Castellanos, 2001; Gelman et al., 2004). Neste trabalho apresentam-se apenas as mais referenciadas.

Considere-se, para a amostra aleatória $X = (X_1, X_2, \ldots, X_n)$, o modelo paramétrico M: $X|\theta \sim f(x|\theta)$, com $\theta \in \Theta$. Suponha-se que foi observada a amostra $x_{obs} = (x_{obs,1}, x_{obs,2}, \ldots, x_{obs,n})$. Seja $T = T(X)$, uma estatística que mede o afastamento do modelo aos dados, de forma que valores elevados de T indiquem uma menor compatibilidade. O valor-p para a estatística T é definido como

$$\Pr^{m(\cdot)}\left[T(X) \geq t(x_{obs})\right],$$

onde $m(\cdot)$ é a distribuição de probabilidade de T, a qual é obtida consoante o método usado para estimar o valor-p.

valor-p clássico

Do lado frequencista, a forma usual para lidar com o vector de parâmetros desco-

nhecido, no cálculo do valor-p, consiste em substitui-lo por uma estimativa de máxima verosimilhança $\hat{\theta}$, uma estimativa de máxima verosimilhança condicional $\hat{\theta}_c$, ou definindo uma estatística suficiente U, tal que a distribuição de T condicional a U não dependa de θ (*e.g.*, Bayarri e Berger, 2000; Robins et al., 2000).

Por exemplo, o valor-p por estimativa directa (*plug in p-value*) é definido por

$$\Pr^{f(\cdot|\hat{\theta}_{obs})}\left[T(X) \geq t(x_{obs})\right]$$

onde $\hat{\theta}_{obs}$ maximiza $f(x_{obs}|\theta)$ e $f(\cdot|\hat{\theta}_{obs})$ é a distribuição de T onde θ é substituído por $\hat{\theta}_{obs}$.

valor-p bayesiano

Os bayesianos têm uma forma natural de "eliminar" os parâmetros que consiste em integrar no correspondente vector de parâmetros. São várias as propostas apresentadas na literatura para o cálculo do valor-p bayesiano, por exemplo, o valor-p preditivo *a priori* de Box (1980), o valor-p preditivo *a posteriori* de Guttman (1967) e Rubin (1984), o valor-p de discrepância de Gelman e Meng (1996) e Meng (1994), o valor-p preditivo condicional e o valor-p preditivo *a posteriori* parcial de Bayarri e Berger (2000).

Box (1980) sugere a utilização da distribuição preditiva *a priori* de T dada por

$$p(t) = \int_{\Theta} f(t|\theta)h(\theta)d\theta,$$

para o cálculo do valor-p, uma vez que esta distribuição não depende de θ, dando origem ao valor-p preditivo *a priori*, que se representa por

$$p_{prior} = \Pr^{p(t)}\left[T(X) \geq t(x_{obs})\right].$$

Segundo Bayarri e Berger (2000), a principal desvantagem do valor-p preditivo *a priori* é a sua dependência na distribuição *a priori* $h(\theta)$. Como a distribuição preditiva *a priori* de T mede a verosimilhança dos dados relativamente ao modelo paramétrico

(informação amostral e distribuição *a priori*), se o modelo for colocado em questão, fica-se na dúvida se esse facto teve como origem a inadequabilidade da distribuição amostral ou da distribuição *a priori*. A utilização de distribuições *a priori* não informativas, para contornar o problema, não se revela uma vantagem inerente, uma vez que estas são, usualmente, impróprias, inviabilizando o próprio cálculo do valor-p preditivo *a priori*.

Para ultrapassar a dependência na distribuição *a priori* $h(\theta)$, no cálculo do valor-p, Guttman (1967) e Rubin (1984) sugerem que se utilize a distribuição preditiva *a posteriori* de T, $p(t|x_{obs}) = \int_\Theta f(t|\theta)h(\theta|x_{obs})d\theta$, dando origem ao valor-$p$ preditivo *a posteriori*

$$p_{post} = \Pr^{p(t|x_{obs})}\left[T(X) \geq t(x_{obs})\right].$$

Gelman et al. (1996) generalizam o valor-p preditivo *a posteriori* permitindo que a estatística T não dependa só dos dados mas também do vector de parâmetros. Para isso, substituem a estatística $T = T(X)$ por $T = T(X, \theta)$, a qual designam por medida de discrepância, permitindo, deste modo, uma comparação mais directa entre os dados e as características populacionais. O valor-p resultante é designado por Robins et al. (2000) de valor-p de discrepância (*discrepancy p-value*).

Uma medida de discrepância muito utilizada para uma validação global do modelo é

$$T(X, \theta) = \sum_{i=1}^{n} \frac{(X_i - E[X_i|\theta])^2}{Var[X_i|\theta]}.$$

Para Bayarri e Berger (2000) as principais vantagens do valor-p preditivo *a posteriori* são: (i) permitir que as distribuições *a priori* não informativas (usualmente impróprias) possam ser utilizadas, uma vez que a distribuição *a posteriori* é, geralmente, própria; (ii) a distribuição preditiva *a posteriori* é mais influenciada pela distribuição amostral do que pela distribuição *a priori* e, consequentemente, o valor-p não é tão sensível a alterações da distribuição *a priori*; e (iii) a facilidade de cálculo via métodos MCMC.

As criticas apontadas, quer por Bayarri e Berger (1999) quer por Robins et al. (2000), à utilização do valor-p preditivo *a posteriori* são: (i) o "duplo uso" dos dados,

em primeiro lugar para a determinação da distribuição *a posteriori* de θ, e em segundo lugar na utilização desta para obter a distribuição preditiva de T; e (ii) em parte devido a (i), o valor-p não ser uniformente distribuído, sendo visto como um método mais conservativo do que o usual.

Para contornar o problema do "duplo uso" dos dados, Bayarri e Berger (1999) sugerem a adopção de um procedimento tipo *validação cruzada*, isto é, parte da amostra observada deve ser utilizada para determinar a distribuição *a posteriori* e a outra parte da amostra observada é utilizada para determinar a distribuição preditiva de T, tendo sempre em consideração as limitações desta técnica, nomeadamente quando a dimensão da amostra é pequena.

Alternativamente, os mesmos autores propõem o valor-p preditivo condicional (*condicional predictive p-value*) e o valor-p preditivo *a posteriori* parcial (*partial posterior predictive p-value*) com o objectivo de manter as vantagens dos valores-p preditivo *a priori* e preditivo *a posteriori*, sem as desvantagens dos mesmos, isto é, o valor-p deve: (i) ser baseado na distribuição preditiva *a priori*; (ii) ser mais influenciado pela distribuição amostral de X do que pela distribuição *a priori* de θ; (iii) permitir o uso de distribuições *a priori* não informativas (impróprias); e (iv) não envolver a utilização do "duplo uso" dos dados.

Tanto o valor-p preditivo condicional, como o valor-p preditivo *a posteriori* parcial, segundo os autores, alcançam os objectivos referidos anteriormente e são assintoticamente uniformes (Robins et al., 2000). No entanto, na maior parte das situações, não são possíveis de determinar analiticamente e, além disso, os métodos MCMC são bastante complicados, em comparação com o processo para a determinação do valor-p preditivo *a posteriori* e do valor-p de discrepância. Portanto, não é usual a sua utilização.

Aspectos Computacionais

Apenas em situações muito simples é possível determinar analiticamente cada valor-p apresentado. A técnica usual para o seu cálculo baseia-se na sua estimação simulando

amostras da distribuição preditiva (*a priori*, *a posteriori* ou condicional) e comparando, à custa da estatística $T(X)$ ou da medida de discrepância $T(X, \theta)$, as amostras simuladas com a amostra observada.

Veja-se, por exemplo, o procedimento a seguir no caso de se pretender estimar o valor-p preditivo *a posteriori* e o valor-p de discrepância.

Seja $X^{rep} = (X_1^{rep}, X_2^{rep}, \ldots, X_n^{rep})$ um vector independente e identicamente distribuído com X, ou seja, uma réplica de X, e

$$p(x^{rep}|x_{obs}) = \int_\Theta f(x^{rep}|\theta)h(\theta|x_{obs})d\theta$$

a correspondente distribuição preditiva *a posteriori* de X^{rep}.

O valor-p preditivo *a posteriori* vem então definido por

$$p_{post}(x_{obs}) = \Pr^{p(x^{rep}|x_{obs})}(T(X^{rep}) \geq t(x_{obs})),$$

isto é, a probabilidade dos dados replicados poderem ser mais extremos que os dados observados, onde $p(x^{rep}|x_{obs})$, na probabilidade, indica que esta é calculada com respeito à distribuição preditiva *a posteriori* de X^{rep} dado x_{obs}. Desta forma, o valor-p é dado por

$$p_{post}(x_{obs}) = \int I_A(x^{rep})p(x^{rep}|x_{obs})dx^{rep},$$

onde $A = \{x^{rep} : t(x^{rep}) \geq t(x_{obs})\}$ e I_A é a função indicatriz do conjunto A, isto é,

$$I_A(x^{rep}) = \begin{cases} 1 & se \quad x^{rep} \in A \\ 0 & se \quad x^{rep} \notin A \end{cases}$$

Consequentemente,

$$p_{post}(x_{obs}) = \int\int I_A(x^{rep})f(x^{rep}|\theta)h(\theta|x_{obs})d\theta dx^{rep}.$$

Assim, para estimar o valor-p preditivo *a posteriori* procede-se de acordo com o seguinte algoritmo:

Algoritmo 1

1. Simula-se uma amostra de dimensão L, $(\theta^{rep,1}, \theta^{rep,2}, \ldots, \theta^{rep,L})$, da distribuição *a posteriori* $h(\theta|x_{obs})$;

2. Para cada $\theta^{rep,l}$, $l = 1, 2, \ldots, L$

   (a) simula-se uma amostra $x^{rep,l}$ da distribuição $f(x|\theta^{rep,l})$;

   (b) calcula-se $t(x_{obs})$ e $t(x^{rep,l})$;

3. Determina-se uma estimativa do valor-p preditivo *a posteriori*, comparando os valores de $t(x^{rep,l})$ com os valores de $t(x_{obs})$ e calculando a proporção das L observações simuladas para as quais $t(x^{rep,l})$ é maior ou igual a $t(x_{obs})$:

$$\hat{p}_{post}(x_{obs}) = \frac{1}{L}\sum_{l=1}^{L} I_A(x^{rep,l}).$$

Pode representar-se graficamente, através de um histograma, os valores simulados de $t(x^{rep,l})$ e identificar, nesse gráfico, o valor de $t(x_{obs})$, permitindo ter uma prespectiva gráfica da estimativa calculada. Se o modelo for adequado, a proporção de valores para as quais $t(x^{rep,l})$ é maior ou igual a $t(x_{obs})$ deve ser elevada.

De forma semelhante, calcula-se o valor-p de discrepância. Seja

$$p(x^{rep}, \theta|x_{obs}) = f(x^{rep}|x_{obs}, \theta)h(\theta|x_{obs})$$

a correspondente distribuição conjunta *a posteriori* de X^{rep} e θ. Uma vez que X^{rep} é independente de X, tem-se que $f(x^{rep}|x_{obs}, \theta) = f(x^{rep}|\theta)$ e, então,

$$p(x^{rep}, \theta|x_{obs}) = f(x^{rep}|\theta)h(\theta|x_{obs}).$$

O valor-p de discrepância é agora definido por

$$p_{dis}(x_{obs}) = \int\int I_B(x^{rep}, \theta)f(x^{rep}|\theta)h(\theta|x_{obs})d\theta dx^{rep},$$

onde $B = \{(x^{rep}, \theta) : t(x^{rep}, \theta) \geq t(x_{obs}, \theta)\}$ e I_B é a função indicatriz do conjunto B.

Para estimar o valor-p de discrepância procede-se de acordo com o seguinte algoritmo:

Algoritmo 2

1. Simula-se L valores, $(\theta^{rep,1}, \theta^{rep,2}, \ldots, \theta^{rep,L})$, da distribuição *a posteriori* $h(\theta|x_{obs})$;
2. Para cada $\theta^{rep,l}$, $l = 1, 2, \ldots, L$
   (a) simula-se uma amostra $x^{rep,l}$ da distribuição $f(x|\theta^{rep,l})$;
   (b) calcula-se $t(x_{obs}, \theta^{rep,l})$ e $t(x^{rep,l}, \theta^{rep,l})$;
3. Determina-se uma estimativa do valor-p de discrepância:

$$\hat{p}_{dis}(x_{obs}) = \frac{1}{L}\sum_{l=1}^{L} I_B(x^{rep,l}, \theta^{rep,l}).$$

Se o modelo for adequado, a proporção de valores para os quais $t(x^{rep,l}, \theta^{rep,l})$ é maior ou igual a $t(x_{obs}, \theta^{rep,l})$ deve estar próximo de 0.5, isto é, o valor ideal do valor-p de discrepância é 0.5. Para Gelman et al. (2004), um modelo é "suspeito" se o valor observado do valor-p de discrepância for inferior a 0.05 ou superior a 0.95, indicando assim que o padrão observado será diferente das réplicas, se o modelo for verdadeiro. Um valor-p extremo implica que o modelo não deve ser utilizado para representar os dados.

Informalmente, a representação gráfica do diagrama de dispersão de $t(x_{obs}, \theta^{rep,l})$ contra $t(x^{rep,l}, \theta^{rep,l})$ ou do histograma das diferenças $t(x_{obs}, \theta^{rep,l}) - t(x^{rep,l}, \theta^{rep,l})$ permitem ter uma perspectiva gráfica da adequabilidade do modelo em estudo. Se o modelo for adequado, a nuvem de pontos acima e abaixo da bissectriz do primeiro quadrante do diagrama de dispersão deve ser idêntica ou, no caso do histograma, este deve incluir a abcissa zero.

Como um modelo pode parecer inadequado devido a diversas razões não controláveis, o valor-p pode ser estimado para diferentes estatísticas, $T(X)$, ou medidas de discrepância, $T(X, \theta)$, de modo a contemplar a validação de várias situações relevantes.

Conforme referem Gelman et al. (2004), "*The relevant goal is not to answer the question, "Do the data come from the assumed model" (to which the answer is almost always no), but to quantify the discrepancies between data and model, and assess whether they could have arisen by chance, under the model's own assumptions.*".

Resumindo, todas as propostas apresentadas para o cálculo do valor-p têm as suas limitações. O valor-p preditivo *a priori*, condicional e *a posteriori* parcial são assintoticamente uniformes. No entanto, o valor-p preditivo *a priori* não está definido quando a distribuição *a priori* é imprópria. O valor-p preditivo condicional e preditivo *a posteriori* parcial são mais complexos e mais difíceis de simular que todos os outros. O valor-p por estimativa directa, preditivo *a posteriori* e de discrepância podem ser muito conservativos, uma vez que no cálculo de cada um deles o conjunto de dados observado é utilizado duas vezes.

3.2 Teste do qui-quadrado

O teste do qui-quadrado, como teste de ajustamento, foi desenvolvido por Pearson (1900) e é um teste que é utilizado para testar a hipótese de que uma determinada amostra aleatória tenha sido extraída de uma população com uma distribuição especificada. Embora existam muitos outros testes de ajustamento clássicos, alguns até mais potentes do que este, o teste de ajustamento do qui-quadrado tem prevalecido no tempo uma vez que é bastante intuitivo, simples e de fácil aplicação.

Embora de índole básica, e apenas por questões formais, apresenta-se seguidamente o teste do qui-quadrado clássico para posteriormente fazer a passagem para a proposta bayesiana.

Teste do qui-quadrado clássico

Considere-se uma amostra $x = (x_1, x_2, \ldots, x_n)$, observação de uma amostra aleatória $X = (X_1, X_2, \ldots, X_n)$, constituída por n observações independentes e identicamente

distribuídas, que está classificada ou que pode ser agrupada em k classes, designadas, por exemplo, por C_j, para $j = 1, 2, \ldots, k$, e seja m_j o número de observações, X_i, que caem em cada uma das classes.

Genericamente, o modelo natural para o número de observações nas diferentes classes é dado pela distribuição multinomial com vector de parâmetros $(p_1, p_2, \ldots, p_k)$, onde $p_j = \Pr[X_i \in C_j]$ e $\sum_{j=1}^{k} p_j = 1$. Consequentemente, o número esperado de observações em cada uma das classes, sob a hipótese da distribuição especificada, é $e_j = np_j$.

A estatística de teste utilizada para validar a distribuição (modelo) compara os valores observados com os valores esperados nas diversas classes. Se a distribuição está completamente especificada, isto é, não depende de parâmetros desconhecidos, a estatística de teste é dada por

$$Q_n = \sum_{j=1}^{k} \frac{(m_j - np_j)^2}{np_j},$$

cuja distribuição assintótica é um qui-quadrado com k-1 graus de liberdade, isto é, $Q_n \overset{\circ}{\sim} \chi^2_{(k-1)}$, conforme foi demonstrado por Pearson (1900).

A aproximação é, em geral, aceite quando todos os valores esperados são superiores a 5; caso contrário, alguns autores optam por agregar classes, não sendo unânime esta decisão entre os estatísticos.

Para um dado nível de significância α, fixo, rejeita-se H_0, isto é, rejeita-se a hipótese da amostra provir de uma população com a distribuição especificada se

$$Q_{n,obs} \geq \chi^2_{(k-1)}(1-\alpha),$$

onde $Q_{n,obs}$ representa o valor observado da estatística Q_n e $\chi^2_{(k-1)}(1-\alpha)$ representa o quantil de probabilidade $(1-\alpha) \times 100\%$ de uma distribuição qui-quadrado com $(k-1)$ graus de liberdade.

No caso da distribuição não estar completamente especificada e, portanto, depender de um vector de parâmetros desconhecido $\theta \in \Theta$, Cramér (1946) provou que a

distribuição assintótica da estatística de teste

$$Q_n(\hat{\theta}) = \sum_{j=1}^{k} \frac{(m_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})}$$

é um qui-quadrado com $(k - s - 1)$ graus de liberdade, $Q_n(\hat{\theta}) \overset{\circ}{\sim} \chi^2_{(k-s-1)}$, onde $\hat{\theta}$ é o vector de parâmetros estimados de dimensão s, utilizando estimadores de máxima verosimilhança.

Para um dado nível de significância α, fixo, rejeita-se H_0, isto é, rejeita-se a hipótese da amostra provir de uma população com a distribuição especificada se

$$Q_{n,obs}(\hat{\theta}) \geq \chi^2_{(k-s-1)}(1 - \alpha),$$

onde $Q_{n,obs}(\hat{\theta})$ representa o valor observado de $Q_n(\hat{\theta})$ e $\chi^2_{(k-s-1)}(1 - \alpha)$ representa o quantil de probabilidade $(1-\alpha)\times 100\%$ de uma distribuição qui-quadrado com $(k-s-1)$ graus de liberdade.

Também se pode calcular o valor-p, neste caso dado por

$$\text{valor-}p = \Pr(Q_n(\hat{\theta}) \geq Q_{n,obs}(\hat{\theta})|H_0),$$

e rejeita-se H_0 se o valor-p for inferior a α.

O teste de ajustamento do qui-quadrado de Pearson, adapta-se facilmente quer a dados discretos, quer a dados contínuos. Como se trata de um teste assintótico, ele deve ser utilizado apenas quando a dimensão da amostra é razoável, não sendo apropriado no caso de amostras pequenas. Relativamente aos dados contínuos, existe sempre a questão da arbitrariedade na escolha das classes para agrupar os dados (intervalos, neste caso), sendo o resultado deste teste altamente influenciado por esse agrupamento.

Teste do qui-quadrado bayesiano

Johnson (2004) apresenta uma alternativa bayesiana ao teste do qui-quadrado clássico, que denomina de teste do qui-quadrado bayesiano. O autor avalia a adequabilidade de uma determinada distribuição de probabilidade utilizando a estatística de

teste de Pearson, mas propondo que o vector de parâmetros desconhecido seja obtido por amostragem da respectiva distribuição *a posteriori*, e prova que a distribuição assintótica desta estatística de teste, para muitos modelos estatísticos, é a distribuição qui-quadrado com $(k-1)$ graus de liberdade, $\chi^2_{(k-1)}$, ou seja, independente da dimensão do vector de parâmetros.

Considere-se as variáveis aleatórias X_i, para $i = 1, 2, \ldots, n$, contínuas, independentes e identicamente distribuídas, com função densidade de probabilidade $f(x_i|\theta)$ condicional a um vector de parâmetros s-dimensional, $\theta \in \Theta \subset \mathbb{R}^s$. É usual utilizar uma distribuição *a priori* não informativa para θ. Seja $F(\cdot|\theta)$ a função de distribuição de X_i, e considere-se que $\breve{\theta}$ é obtido por amostragem a partir da sua distribuição *a posteriori*, isto é, é um valor gerado de $h(\theta|x)$.

Seguidamente, o intervalo de variação da função de distribuição, $[0, 1]$, é dividido em k sub-intervalos de igual amplitude, isto é, $0 \equiv a_0 < a_1 < \ldots < a_{k-1} < a_k \equiv 1$, com $p_j = a_j - a_{j-1} = \frac{1}{k}$, para $j = 1, \ldots, k$, por forma a construir as classes equiprováveis.

Relativamente ao número de sub-intervalos a definir, Mann e Wald (1942) sugerem a utilização de $3.8(n-1)^{0.4}$ sub-intervalos, onde n é a dimensão da amostra. No entanto, muitos autores consideram que este critério origina muitas classes e, consequentemente, uma perda de potência do teste. Tendo por base estudos de simulação, nomeadamente para o estudo da adequabilidade de modelos contínuos, em particular modelos normais, Koehler e Gan (1990) propõem a utilização do critério de Mann e Hald dividido por 4. Johnson (2004) opta por definir para o número de sub-intervalos um valor aproximado de $n^{0.4}$.

Finalmente, cada uma das observações amostrais, x_i, é alocada a uma e só uma classe de acordo com o valor observado da função de distribuição acumulada condicional ao vector de parâmetros $\breve{\theta}$, $F(x_i|\breve{\theta})$, para $i = 1, 2, \ldots, n$.

Considere-se assim que $z_i(\breve{\theta})$ é um vector de dimensão k cujo j-ésimo elemento é definido por

$$z_{i,j} = \begin{cases} 1 & \text{se } F(x_i|\check{\theta}) \in (a_{j-1}, a_j] \\ 0 & \text{em caso contrário} \end{cases}.$$

Finalmente, define-se o vector

$$m(\check{\theta}) = \sum_{i=1}^{n} z_i(\check{\theta}),$$

onde o j-ésimo elemento de $m(\check{\theta})$, $m_j(\check{\theta})$, representa o número de observações que caem na j-ésima classe.

Johnson (2004), tendo por base os trabalhos de Chernoff e Lehmann (1954), Cramér (1946) e Chen (1985), apresenta e prova que

$$Q_n^B(\check{\theta}) = \sum_{j=1}^{k} \frac{(m_j(\check{\theta}) - np_j)^2}{np_j}, \tag{3.2}$$

tem uma distribuição assintótica de um $\chi^2_{(k-1)}$, independente da dimensão do vector de parâmetros θ.

O autor apresenta uma justificação para a diferença do número de graus de liberdade entre este teste e o teste do qui-quadrado clássico. Os s graus de liberdade perdidos ao substituir os s parâmetros desconhecidos por estimadores de máxima verosimilhança no teste do qui-quadrado clássico, são totalmente recuperados quando se substitui o vector s-dimensional θ por valores simulados da respectiva distribuição *a posteriori*.

Este teste de ajustamento pode também ser utilizado para variáveis discretas. Johnson (2004) apresenta duas alternativas. A primeira alternativa, e a mais directa, é proceder exactamente como no caso contínuo. A segunda alternativa consiste em utilizar um procedimento semelhante ao teste do qui-quadrado clássico, isto é, construir as classes de acordo com os valores possíveis da variável aleatória em estudo, seguindo-se o cálculo das probabilidades associadas a cada uma das classes, utilizando valores gerados da distribuição *a posteriori* para o vector de parâmetros.

Seja $f(x_i|\theta)$ a função de probabilidade e

$$p_j(\check{\theta}) = \sum_{i=1}^{n} f(x_i|\check{\theta}) I_{C_j}(x_i).$$

A estatística definida em (3.2) é substituída por

$$Q_n^B(\check{\theta}) = \sum_{j=1}^{k} \frac{(m_j - np_j(\check{\theta}))^2}{np_j(\check{\theta})},$$

que, segundo o referido autor, mantém a distribuição assintótica de um $\chi^2_{(k-1)}$.

A decisão de rejeitar a hipótese H_0 é, para um dado nível de significância α, fixo, dada por

$$Q^B_{n,obs}(\check{\theta}) \geq \chi^2_{(k-1)}(1-\alpha),$$

onde $Q^B_{n,obs}(\check{\theta})$ representa o valor observado da estatística $Q_n^B(\check{\theta})$ e $\chi^2_{(k-1)}(1-\alpha)$ representa o quantil de probabilidade $(1-\alpha)\times 100\%$ de uma distribuição qui-quadrado com $(k-1)$ graus de liberdade.

Alternativamente, Johnson (2004) propõe como regra de decisão, para rejeitar a hipótese nula, a proporção de valores para os quais $Q^B_{n,obs}(\check{\theta})$, utilizando L valores de $\check{\theta}$ gerados da distribuição *a posteriori*, é maior ou igual a um determinado valor crítico (por exemplo, $\chi^2_{(k-1)}(0.95)$). A hipótese nula é rejeitada se a proporção calculada for superior a um valor de corte fixo (*threshold*). Qualquer excesso da proporção calculada, pode ser devido à dependência entre os valores $\check{\theta}$ gerados ou simplesmente porque a distribuição em estudo não se adequa aos dados.

O principal problema com esta regra, reside no facto de não existir qualquer resultado teórico sobre o valor a atribuir ao *threshold*, para um determinado nível de significância, fixo.

3.3 Testes de ajustamento bayesianos

Um teste de ajustamento bayesiano, ao contrário do que sucede na estatística clássica, requer a especificação de um modelo (distribuição) alternativo(a). O modelo alternativo deve ser visto como um modelo mais geral do que o modelo paramétrico, no sentido de poder providenciar um melhor ajustamento. A classe de modelos não paramétricos surge assim como uma possível resposta ao problema apresentado.

A abordagem bayesiana não paramétrica para este tipo de teste, consiste em incorporar (*embed*), de alguma forma, o modelo paramétrico directamente no modelo não paramétrico (ou alternativo). Os dois modelos são seguidamente comparados, utilizando critérios de comparação de modelos, como por exemplo, utilizando o factor de Bayes e, finalmente, um deles é seleccionado. Designa-se este teste por teste de ajustamento bayesiano não paramétrico.

A literatura estatística bayesiana que aborda o teste de ajustamento bayesiano não paramétrico é ainda muito escassa. Veja-se, por exemplo, os artigos pioneiros de Carota e Parmigiani (1996) e Florens et al. (1996), onde os autores fazem uma primeira abordagem ao tema apontando as suas dificuldades e limitações práticas. Conigliani et al. (2000) propõem dois novos testes de ajustamento para dados discretos que são considerados, pelos autores, como alternativas bayesianas ao teste de ajustamento do qui-quadrado clássico. O primeiro teste compara o modelo paramétrico em estudo com o modelo multinomial simples. O segundo teste utiliza o conceito definido no teste de ajustamento bayesiano não paramétrico, isto é, incorpora o modelo paramétrico em estudo no modelo multinomial. O estudo dos referidos autores é apenas desenvolvido para dados discretos e para quando são definidas distribuições *a priori* não informativas (em particular, impróprias) para o vector de parâmetros.

Para dados contínuos, Verdinelli e Wasserman (1998), Berger e Guglielmi (2001) e Tokdar e Martin (2011) apresentam, respectivamente, três novos testes de ajustamento bayesianos não paramétricos, mas apenas associados ao estudo de adequabilidade da

distribuição normal. O modelo não paramétrico é definido por uma mistura de processos gaussianos, uma mistura de árvores de Pólya e uma mistura por processo de Dirichlet, respectivamente. No entanto, o teste de Berger e Guglielmi (2001) é computacionalmente mais acessível e teoricamente mais intuitivo, comparativamente com os outros dois. Além disso, é interessante estudar como este teste pode ser adaptado para o estudo da adequabilidade de outras distribuições, além da distribuição normal.

Na secção seguinte, apresenta-se, resumidamente, a abordagem proposta por Conigliani et al. (2000) para dados discretos. Depois, apresenta-se a abordagem de Berger e Guglielmi (2001), para dados contínuos. No entanto, uma leitura cuidadosa dos respectivos artigos é fundamental para entender muitas das opções dos autores.

3.3.1 Dados discretos

Seja $X = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória, referente a dados discretos, constituída por n observações independentes e identicamente distribuidas, onde cada observação pode ser classificada em uma de $(k+1)$ classes, definidas por C_j, para $j = 0, 1, \ldots, k$, e seja $(m_0, m_1, \ldots, m_k)$ o vector que contém o número de observações que caem em cada uma das $(k+1)$ classes.

Considere-se também que $\mathrm{M}_k(n; p)$ representa a distribuição multinomial de parâmetros n (fixo) e $p = (p_0, p_1, \ldots, p_k)$, com $p_j \geq 0$ e $\sum_{j=0}^{k} p_j = 1$, realçando a dimensionalidade k desta distribuição, uma vez que, por exemplo, $m_k = n - (m_0 + m_1 + \cdots + m_{k-1})$ e $p_k = 1 - (p_0 + p_1 + \cdots + p_{k-1})$.

A abordagem inicial de Conigliani et al. (2000), ou seja, o primeiro teste de ajustamento bayesiano, define dois modelos competitivos para a amostra agrupada em $(k+1)$ classes. O primeiro modelo, ou modelo M_1, pressupõe que os p_j, para $j = 0, 1, \ldots, k$, pertencem a uma determinada família paramétrica $\mathcal{F} = \{p_j(\theta), \theta \in \mathbb{R}^s, s < \infty\}$ caracterizada por $p_j(\theta)$, que depende de um vector de parâmetros s-dimensional θ. Para o segundo modelo, M_2 (o modelo não paramétrico ou alternativo), não é feita qualquer

suposição para os p_j além das suposições usuais, $p_j \geq 0$ e $\sum_{j=0}^{k} p_j = 1$.

Os dois modelos a comparar podem ser representados por:

$$M_1 : \Pr(X_i \in C_j|\theta) = p_j(\theta) \quad \text{e} \quad M_2 : \Pr(X_i \in C_j|p) = p_j, \quad \text{para } j = 0, 1, \ldots, k$$

e, seguidamente, um dos dois modelos é seleccionado utilizando o factor de Bayes fraccionário.

Sob o modelo M_1, a função de verosimilhança é dada por

$$f_1(x|\theta) = \prod_{j=0}^{k} [p_j(\theta)]^{m_j},$$

onde o vector de parâmetros θ segue uma distribuição *a priori* $h_1(\theta)$. Para o modelo alternativo M_2, a função de verosimilhança é dada por

$$f_2(x|p) = \prod_{j=0}^{k} p_j^{m_j},$$

onde o vector de parâmetros $(p_0, p_1, \ldots, p_k)$ tem como distribuição *a priori* a distribuição conjugada natural do modelo amostral, ou seja, a distribuição Dirichlet com hiperparâmetros $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_k) \in \mathbb{R}_+^{k+1}$ e cuja função densidade no simplex k-dimensional

$$S_k = \{p = (p_0, p_1, \ldots, p_{k-1}) : p_j \geq 0, \sum_{j=0}^{k-1} p_j < 1\}$$

é dada por

$$h_2(p) = \frac{\Gamma(\sum_{j=0}^{k} \alpha_j)}{\prod_{j=0}^{k} \Gamma(\alpha_j)} p_0^{\alpha_0 - 1} p_1^{\alpha_1 - 1} \cdots \left(1 - \sum_{j=0}^{k-1} p_j\right)^{\alpha_k - 1}. \tag{3.3}$$

Fazendo o hiperparâmetro α_j, para $j = 0, 1, \ldots, k$, tender para zero, é obtida, no limite, uma distribuição *a priori* imprópria:

$$h_2(p) \propto \prod_{j=0}^{k} p_j^{-1}. \tag{3.4}$$

O cálculo do factor de Bayes fraccionário definido em (2.3), assumindo, inicialmente, a distribuição *a priori* dada em (3.3), tem as seguintes expressões para $p_j(x; b)$, com

$j = 1, 2$:

$$p_1(x; b) = \frac{\int f_1(x|\theta) h_1(\theta) d\theta}{\int f_1(x|\theta)^b h_1(\theta) d\theta} = \frac{\int \prod_{j=0}^{k} p_j(\theta)^{m_j} h_1(\theta) d\theta}{\int \prod_{j=0}^{k} p_j(\theta)^{b m_j} h_1(\theta) d\theta} \tag{3.5}$$

e

$$\begin{aligned} p_2(x; b) &= \frac{\int_{S_k} f_2(x|p) h_2(p) dp}{\int_{S_k} f_2(x|p)^b h_2(p) dp} \\ &= \frac{\int_{S_k} \prod_{j=0}^{k} p_j^{m_j+\alpha_j-1} dp}{\int_{S_k} \prod_{j=0}^{k} p_j^{b m_j+\alpha_j-1} dp} \\ &= \frac{\Gamma(bn+c)}{\Gamma(n+c)} \prod_{j=0}^{k} \frac{\Gamma(m_j+\alpha_j)}{\Gamma(b m_j+\alpha_j)} \end{aligned} \quad , \tag{3.6}$$

onde $c = \sum_{j=0}^{k} \alpha_j$. À medida que cada $\alpha_j \to 0$, tem-se, no limite, que (3.6) é substituída por

$$p_2(x; b) = \frac{\Gamma(bn)}{\Gamma(n)} \prod_{j=0,\ m_j>0}^{k} \frac{\Gamma(m_j)}{\Gamma(b m_j)} \ . \tag{3.7}$$

Falta apenas definir o valor a atribuir a $b = n_1/n$, isto é, o valor para a proporção do conjunto de dados, que vai ser utilizada para a amostra de treino. De acordo com estudos de simulação feitos em O'Hagan (1995) e em Conigliani e O'Hagan (1996), sobre os possíveis valores para b, os autores optam por utilizar $n_1 = (k+1)$, ou seja, atribuir à dimensão da amostra mínima uma observação por classe. No caso do número de classes ser infinito, escolhem $n_1 = s_1 + 1$, onde s_1 é tal que para todo o $j > s_1$ tem-se que $m_j = 0$.

Comparando os pressupostos deste teste de ajustamento com o teste do qui-quadrado clássico, pode desde já referir-se que algumas das limitações deste último não se verificam, nomeadamente: (i) nenhum dos valores para $p_j(x; b)$, para $j = 1, 2$, é baseado em aproximações; e (ii) ambos os valores de $p_j(x; b)$ estão bem definidos, mesmo que alguma das classes não tenha observações, ou tenha poucas observações.

Conigliani et al. (2000) aplicam o factor de Bayes fraccionário a dois modelos de dados discretos, nomeadamente, aos modelos cujas distribuições amostrais são as distribuições binomial e Poisson, respectivamente. A metodologia apresentada é utilizada pelos referidos autores em alguns exemplos reais e fictícios. Os resultados obtidos, para esses exemplos, permitem concluir que o factor de Bayes fraccionário fornece conclusões

sensatas quando o número de classes não é muito grande. No entanto, quando o número de classes é grande, como para o modelo Poisson, os resultados obtidos podem ser menos significativos que os obtidos por outros métodos de estudo de adequabilidade, como por exemplo, o valor-p de discrepância ou mesmo outros métodos de estudo de adequabilidade clássicos referidos no artigo.

Uma explicação para esta situação, pode ter a ver com a diferença entre o número de parâmetros do modelo M_1, usualmente um número finito e pequeno, e o número de parâmetros do modelo M_2, que no mínimo é igual ao número de classes definidas.

Para contornar este facto, Conigliani et al. (2000) propõem uma estrutura hierárquica na distribuição *a priori* do modelo alternativo. Conforme referem os autores, "*To overcome this difficulty, and reduce the number of parameters in M_2, we replace the alternative models by a hierarchical one, constructed by embedding the parametric model in a nonparametric model. ... which was found by Carota et al. (1996) to work well in a wide range of discrete data problems*".

A estrutura hierárquica na distribuição *a priori* do modelo alternativo é, então, definida por: no primeiro nível, $p = (p_0, p_1, \ldots, p_{k-1})$ segue uma distribuição *a priori* (própria) conjugada natural do modelo amostral, a distribuição Dirichlet definida em (3.3); no segundo nível, os autores pressupõem que a média *a priori* de cada p_j pertence à família paramétrica $\mathcal{F}$, isto é

$$E[p_j] = \frac{\alpha_j}{\sum_{j=0}^{k} \alpha_j} = \frac{\alpha_j}{c} = p_j(\theta), \text{ para } j = 0, 1, \ldots, k, \tag{3.8}$$

onde o vector de parâmetros θ tem distribuição *a priori* não informativa, $h_1(\theta)$, ou seja, mantém a distribuição *a priori* do modelo M_1.

De (3.8) obtém-se que

$$\alpha_j = c p_j(\theta)$$

e, consequentemente,

$$Var[p_j] = \frac{p_j(\theta)(1 - p_j(\theta))}{c + 1},$$

onde c é denominado de parâmetro de concentração, que é considerado fixo. Os autores afirmam que "*... c can be seen as representing a belief about how "close" the true distribution should be to a member of the parametric family, if in fact the alternative model holds. As c increases, the prior variances of the p_js decrease and the p_js become closer to the $p_j(\theta)$s. In the limit, as c goes to infinity, the alternative model, M_2, coincides with M_1.*". Relembrando os conceitos introduzidos na secção 2.2.1, o modelo alternativo apresentado é um modelo bayesiano não paramétrico e o teste é um teste de ajustamento bayesiano não paramétrico.

Com base em todos estes novos pressupostos, o cálculo do factor de Bayes fraccionário mantém a expressão de $p_1(x;b)$ dada em (3.5). No entanto, a expressão de $p_2(x;b)$, apresenta as seguintes alterações:

A estrutura hierárquica na distribuição *a priori*, faz com que a nova distribuição *a priori* seja definida por

$$h_2(p) = \int h_2(p|\theta)h_1(\theta)d\theta$$

e, consequentemente, de (2.4),

$$\begin{aligned} p_2(x;b) &= \frac{\int_{S_k} f_2(x|p) \int h_2(p|\theta)h_1(\theta)d\theta\, dp}{\int_{S_k} f_2(x|p)^b \int h_2(p|\theta)h_1(\theta)d\theta\, dp} \\ &= \frac{\int_{S_k} \prod_{j=0}^k p_j^{m_j} \int [B(\alpha)]^{-1} \prod_{j=0}^k p_j^{cp_j(\theta)-1} h_1(\theta)d\theta\, dp}{\int_{S_k} \prod_{j=0}^k p_j^{bm_j} \int [B(\alpha)]^{-1} \prod_{j=0}^k p_j^{cp_j(\theta)-1} h_1(\theta)d\theta\, dp} \\ &= \frac{\int \int_{S_k} \prod_{j=0}^k p_j^{m_j+cp_j(\theta)-1} dp\, [B(\alpha)]^{-1} h_1(\theta)d\theta}{\int \int_{S_k} \prod_{j=0}^k p_j^{bm_j+cp_j(\theta)-1} dp\, [B(\alpha)]^{-1} h_1(\theta)d\theta} \end{aligned},$$

onde $B(\alpha) = \dfrac{\sum_{j=0}^k \Gamma(cp_j(\theta))}{\Gamma(c)}$.

Como $\displaystyle\int_{S_k} \prod_{j=0}^k p_j^{m_j+cp_j(\theta)-1} dp$ define o integral de Dirichlet, tem-se

$$\begin{aligned} \int_{S_k} \prod_{j=0}^k p_j^{m_j+cp_j(\theta)-1} dp &= \frac{\prod_{j=0}^k \Gamma\{m_j + cp_j(\theta)\}}{\Gamma\{\sum_{j=0}^k (m_j + cp_j(\theta))\}} \\ &= \frac{\prod_{j=0}^k \Gamma\{m_j + cp_j(\theta)\}}{\Gamma(n+c)} \end{aligned}.$$

O mesmo raciocínio é aplicado ao integral de Dirichlet do denominador. Desta forma,

$$p_2(x;b) = \frac{\Gamma(bn+c)}{\Gamma(n+c)} \frac{\int \prod_{j=0}^{k} \frac{\Gamma\{m_j + cp_j(\theta)\}}{\Gamma\{cp_j(\theta)\}} h_1(\theta) d\theta}{\int \prod_{j=0}^{k} \frac{\Gamma\{bm_j + cp_j(\theta)\}}{\Gamma\{cp_j(\theta)\}} h_1(\theta) d\theta} . \tag{3.9}$$

Quando é considerada uma estrutura hierárquica na distribuição *a priori* do modelo alternativo e segundo Conigliani et al. (2000), o valor atribuído à dimensão da amostra mínima é 1. Consequentemente, tem-se que, neste caso, $b = 1/n$.

O factor de Bayes fraccionário com estrutura hierárquica depende do parâmetro de concentração c. Consequentemente, é necessário atribuir valores a esse parâmetro. Para melhor compreender como os diferentes valores atribuídos ao parâmetro c influenciam o correspondente factor de Bayes fraccionário hierárquico, apresentam-se seguidamente alguns exemplos de aplicação.

Conigliani et al. (2000) utilizam o factor de Bayes fraccionário para contornar o problema das distribuições *a priori* não informativas impróprias, usuais neste tipo de problema. No entanto, no Capítulo 2, secção 2.3, foram apresentados outros critérios de comparação de modelos, nomeadamente o factor pseudo-Bayes. Será que existem diferenças significativas no desempenho destes diferentes factores de Bayes? Através de alguns exemplos de aplicação apresentam-se os diferentes factores de Bayes e estudam-se os resultados obtidos.

3.3.2 Exemplos de aplicação

Nesta secção, exemplifica-se como se aplicam alguns dos métodos de estudo da adequabilidade do modelo Poisson, nomeadamente, o factor de Bayes fraccionário simples, o factor de Bayes fraccionário hierárquico, o valor-p de discrepância e o factor pseudo-Bayes. Este último método é uma nova proposta em estudo que utiliza a teoria desen-

volvida para o factor de Bayes fraccionário, sem estrutura hierárquica, mas substituí o cálculo do factor de Bayes fraccionário pelo cálculo do factor pseudo-Bayes, definido em (2.5). Tem-se como principal objectivo comparar os valores dos factores obtidos pelos diferentes métodos e tentar chegar a algumas conclusões úteis.

A distribuição de Poisson é usualmente utilizada para descrever fenómenos aleatórios que envolvam a contagem de acontecimentos raros que ocorrem em determinado período de tempo ou espaço. A igualdade da média e da variância é uma característica importante da distribuição de Poisson, denominada usualmente por *equal dispersion*. Esta característica, no entanto, dificulta a aplicação da distribuição de Poisson a dados discretos que têm, ou variância superior à média, denominada por *over dispersion*, ou variância inferior à média, denominada por *under dispersion*. Existem, contudo, outras distribuições alternativas à distribuição de Poisson, apropriadas este tipo de dados, mas que permitem controlar a dispersão, nomeadamente, a distribuição binomial negativa e a distribuição binomial.

O modelo Poisson define-se para observações $x_1, x_2, \ldots, x_n$ independentes, condicionalmente a θ, simbolicamente dado por

$$X_i|\theta \overset{\text{iid}}{\sim} \text{Po}(\theta), \quad \text{para } i = 1, 2, \ldots, n \text{ e } \theta > 0,$$

e cuja função de probabilidade é dada por

$$f(x_i|\theta) = \frac{e^{-\theta}\theta^{x_i}}{x_i!}, \text{ para } x_i = 0, 1, \ldots.$$

Neste modelo e dada a natureza dos x_i, as classes C_j identificam-se com cada valor de x_i, sendo a última classe definida de acordo com as observações. Assim, para definir o factor de Bayes fraccionário, proposto por Conigliani et al. (2000) e apresentado na secção 3.3.1, a distribuição de Poisson é a família paramétrica caracterizada por $p_j(\theta)$, para $j = 0, 1, \ldots$, que vai representar o modelo M_1. Desta forma, tem-se que

$$p_j(\theta) = \frac{e^{-\theta}\theta^{j}}{j!}.$$

A distribuição *a priori* conjugada natural para θ é a distribuição gama com parâmetros a e b, simbolicamente $\theta \sim \text{Ga}(a, b)$.

A distribuição *a priori* não informativa (imprópria) para θ e associada ao modelo M_1, é dada por

$$h_1(\theta) \propto \theta^{-1}, \tag{3.10}$$

de acordo com o critério de Haldane e é dada por

$$h_1(\theta) \propto \theta^{-1/2},$$

de acordo com o critério de Jeffreys.

Para exemplificar os métodos de estudo da adequabilidade do modelo Poisson, referidos anteriormente, utilizam-se amostras simuladas de três distribuições associadas a dados discretos, nomeadamente, a distribuição de Poisson, a distribuição binomial e a distribuição binomial negativa, sumariadas na Tabela 3.1. Os parâmetros das amostras simuladas, das três distribuições, são escolhidos por forma a que tenham todas a mesma média e cujas variâncias, nomeadamente das distribuições binomial e binomial negativa, estejam próximas e afastadas da respectiva média.

Tabela 3.1: Três distribuições associadas a dados discretos e respectivos parâmetros.

Distribuição	Notação	$f(x_i\|\theta)$	$E(X_i)$	$Var(X_i)$	Suporte
Poisson	$\text{Po}(\theta)$	$\frac{e^{-\theta}\theta^{x_i}}{x_i!}$	θ	θ	$x_i = 0, 1, \ldots$
Binomial	$\text{Bi}(n, \theta)$	$C_{x_i}^{n}\theta^{x_i}(1-\theta)^{n-x_i}$	$n\theta$	$n\theta(1-\theta)$	$x_i = 0, 1, \ldots, n$
Binomial Negativa	$\text{BiN}(r, \theta)$	$C_{x_i}^{x_i+r-1}\theta^{r}(1-\theta)^{x_i}$	$\frac{r(1-\theta)}{\theta}$	$\frac{r(1-\theta)}{\theta^2}$	$x_i = 0, 1, \ldots$

Definem-se, seguidamente, cada uma das expressões práticas para o cálculo dos diferentes factores de Bayes e para o cálculo da estimativa do valor-p de discrepância, para o estudo da adequabilidade da distribuição de Poisson a um conjunto de observações amostrais.

Seja $x = (x_1, x_2, \ldots, x_n)$ a concretização de n variáveis aleatórias independentes e identicamente distribuídas a X_i, referente a dados discretos, onde cada observação é classificada em uma de $(k+1)$ classes, e seja $(m_0, m_1, \ldots, m_k)$ o vector que contém o número de observações x_i iguais a j, para $j = 0, 1, \ldots, k$ e $i = 1, 2, \ldots, n$. Sejam $m = \sum_{j=0}^{k} j m_j$, $n = \sum_{j=0}^{k} m_j$ e b o valor que representa a proporção do conjunto de observações a utilizar para o cálculo do factor de Bayes fraccionário. Considere-se ainda que são utilizadas as distribuições *a priori* não informativas para θ e para $(p_0, p_1, \ldots, p_k)$, definidas em (3.10) e (3.4), respectivamente.

1. O factor de Bayes fraccionário definido em (2.3), utilizando as expressões dadas em (3.5) e (3.7), é calculado como

$$\mathrm{BF}^{21}_{frac}(x; b) = \frac{\dfrac{\Gamma(bn)}{\Gamma(n)} \displaystyle\prod_{j=0,\, m_j>0}^{k} \frac{\Gamma(m_j)}{\Gamma(bm_j)}}{\dfrac{(bn)^{bm}}{n^m} \dfrac{\Gamma(m)}{\Gamma(bm)} \displaystyle\prod_{j=0}^{k} (j!)^{m_j(b-1)}},$$

   onde $b = (k+1)/n$.

2. O factor de Bayes fraccionário hierárquico, mantém a expressão dada em (3.5) e substituí a expressão (3.7) pela expressão dada em (3.9), obtendo-se

$$\mathrm{BF}^{21}_{frac_h}(x; b) = \frac{\dfrac{\displaystyle\int_0^{\infty} \prod_{j=0, m_j>0}^{k} \frac{\Gamma\{m_j + c\frac{e^{-\theta}\theta^j}{j!}\}}{\Gamma\{c\frac{e^{-\theta}\theta^j}{j!}\}} \theta^{-1} d\theta}{\displaystyle\int_0^{\infty} \prod_{j=0, m_j>0}^{k} \frac{\Gamma\{bm_j + c\frac{e^{-\theta}\theta^j}{j!}\}}{\Gamma\{c\frac{e^{-\theta}\theta^j}{j!}\}} \theta^{-1} d\theta}}{\dfrac{(bn)^{bm}}{n^m} \dfrac{\Gamma(n+c)}{\Gamma(bn+c)} \dfrac{\Gamma(m)}{\Gamma(bm)} \displaystyle\prod_{j=0}^{k} (j!)^{m_j(b-1)}},$$

   onde $b = 1/n$ e c é o parâmetro de concentração (fixo) que vai tomar diferentes valores, nomeadamente, $c = 2$, 5, 10, 50, 100 e 200.

3. A nova proposta, o factor pseudo-Bayes a favor do modelo alternativo sem estrutura hierárquica (M_2) e contra o modelo em estudo (M_1) é, por definição, o

quociente entre o produto das ordenadas preditivas condicionais para os respectivos modelos, onde

$$\mathrm{CPO}_j(x_i) = p_j(x_i|x_{(-i)}) = p_j(x)/p_j(x_{(-i)}),$$

para $j = 1, 2$.

Neste caso, não é necessário utilizar métodos de simulação Monte Carlo para determinar as correspondentes CPO, uma vez que as distribuições preditivas *a priori*, $p_j(x)$, para $j = 1, 2$, podem ser obtidas analiticamente. A distribuição preditiva *a priori* para M_1 é dada por

$$p_1(x) = \frac{1}{\prod_{j=0}^{k}(\Gamma(j+1))^{m_j}} \frac{\Gamma(m)}{n^m}. \tag{3.11}$$

A distribuição preditiva *a priori* para M_2 é dada por

$$p_2(x) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=0}^{k} \frac{\Gamma(m_j+\alpha_j)}{\Gamma(\alpha_j)}$$

onde $\alpha = \sum_{j=0}^{k} \alpha_j$. É utilizada, neste caso, a distribuição *a priori* conjugada não informativa de Jeffreys para $(p_0, p_1, \ldots, p_k)$, fazendo os parâmetros da distribuição Dirichlet iguais a 0.5, isto é, $\alpha_j = 0.5$, para $j = 0, 1, \ldots, k$.

Finalmente, para o cálculo da estimativa do valor-p de discrepância, utiliza-se como medida de discrepância

$$T(X, \theta) = \sum_{i=1}^{n} \frac{(X_i - E[X_i|\theta])^2}{Var[X_i|\theta]}.$$

O algoritmo 2, definido na secção 3.1, é executado considerando $L = 10000$ amostras replicadas. É utilizada a distribuição *a priori* não informativa (imprópria) definida em (3.10), uma vez que a correspondente distribuição *a posteriori*, $\mathrm{Ga}(m, n)$, é própria.

Resultados e discussão

Nas Tabelas 3.2 e 3.3 encontram-se os resultados obtidos para 14 amostras simuladas, constituídas por observações independentes e identicamente distribuídas, Po(1), Bi(2,0.5), BiN(1,0.5), Bi(4,0.5), BiN(2,2/3), Bi(10,0.1), BiN(9,0.9), Po(5), Bi(10,0.5), BiN(5,0.5), Bi(20,0.25), BiN(10,2/3), Bi(50,0.1) e BiN(45,0.9), cada amostra com dimensão $n = 100$.

Nas Figuras 3.1 e 3.2 ilustra-se, para duas das amostras, o modo como se obtêm as estimativas do valor-p de discrepância. Ou seja, estas estimativas são obtidas pela proporção de pontos que caem acima da bissectriz do primeiro quadrante.

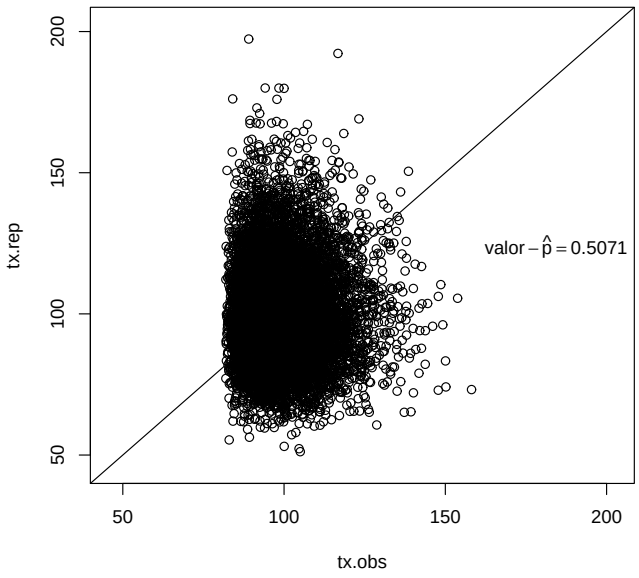

Figura 3.1: Diagrama de dispersão de $t(x^{rep,l}, \theta^{rep,l})$ (ordenadas) *versus* $t(x_{obs}, \theta^{rep,l})$ (abcissas), obtido com os dados de uma amostra de dimensão $n = 100$, simulada de uma distribuição de Poisson, Po(1).

Quando a amostra é simulada de uma distribuição de Poisson, Po(1) e Po(5), o modelo Poisson é validado por qualquer um dos métodos. No entanto, os resultados obtidos quando se utiliza o factor de Bayes fraccionário hierárquico são dependentes

Tabela 3.2: Cálculo dos diferentes factores de Bayes e valor-p de discrepância, no modelo Poisson, para amostras simuladas de várias distribuições.

	Distribuição						
	Po(1)	Bi(2,0.5)	BiN(1,0.5)	Bi(4,0.25)	BiN(2,2/3)	Bi(10,0.1)	BiN(9,0.9)
Amostra	x_1	x_2	x_3	x_4	x_5	x_6	x_7
$\bar{x}$	1.06	1.01	1.14	1.03	0.96	1.01	1.09
s^2	1.05	0.51	1.93	0.68	1.41	0.90	1.27
$\mathrm{BF}^{21}_{frac}(x;b)$	0.0157	61312.77	205.69	0.5708	0.1804	0.1112	0.1363
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=2$	0.0100	51329.50	6.3817	0.9079	0.1000	0.0393	0.0680
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=5$	0.0339	40055.70	8.7044	2.7942	0.3412	0.1048	0.2113
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=10$	0.0839	21918.97	9.4954	5.7054	0.7365	0.2199	0.4772
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=50$	0.4597	852.78	10.1675	9.7543	2.7841	0.8772	1.7219
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=100$	0.7056	126.98	8.7613	6.7839	3.3791	1.1739	1.9744
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=200$	0.9042	22.59	6.7255	4.0124	3.1559	1.3246	1.8550
$\mathrm{BF}^{21}_{pseudo}(x)$	0.1472	185306.20	437.45	7.1549	4.2529	0.5299	1.3538
valor-$\hat{p}$	0.5071	0.9987	0.0026	0.9765	0.0208	0.6888	0.1927

Tabela 3.3: Cálculo dos diferentes factores de Bayes e valor-p de discrepância, no modelo Poisson, para amostras simuladas de várias distribuições (continuação).

	Distribuição						
	Po(5)	Bi(10,0.5)	BiN(5,0.5)	Bi(20,0.25)	BiN(10,2/3)	Bi(50,0.1)	BiN(45,0.9)
Amostra	x_8	x_9	x_10	x_11	x_12	x_13	x_14
$\bar{x}$	5.06	5.02	5.19	5.45	4.4	5.02	5.16
s^2	5.43	2.93	10.46	3.30	6.9	4.02	5.63
Método							
$\mathrm{BF}^{21}_{frac}(x;b)$	0.11×10^{-3}	2.8992	35.9815	0.5350	1.6678	0.0036	0.0028
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=2$	0.77×10^{-8}	0.0480	2.13×10^{-5}	0.0014	2.52×10^{-5}	0.16×10^{-5}	0.82×10^{-6}
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=5$	0.85×10^{-6}	0.8283	0.0037	0.1022	0.0015	0.0001	0.45×10^{-4}
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=10$	0.28×10^{-4}	3.9584	0.2036	1.6207	0.0212	0.0023	0.0008
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=50$	0.0109	5.1681	84.6222	31.4954	0.3970	0.1104	0.0780
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=100$	0.0358	1.5996	115.4283	25.1313	1.0862	0.1437	0.1698
$\mathrm{BF}^{21}_{frac_h}(x;b)$, $c=200$	0.0601	0.4499	51.2552	12.8834	1.6325	0.1139	0.2247
$\mathrm{BF}^{21}_{pseudo}(x)$	0.0010	31.6366	1021.11	2.6699	12.1462	0.0081	0.0382
valor-$\hat{p}$	0.3101	0.9994	0	0.9979	0.0012	0.9108	0.2663

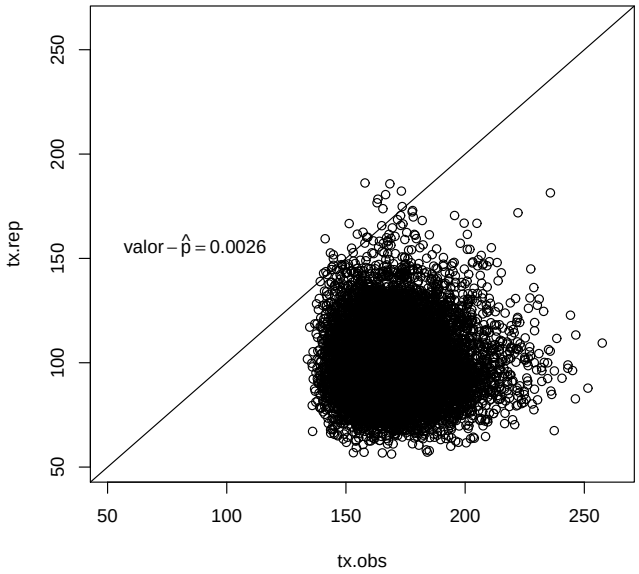

Figura 3.2: Diagrama de dispersão de $t(x^{rep,l}, \theta^{rep,l})$ (ordenadas) *versus* $t(x_{obs}, \theta^{rep,l})$ (abcissas), obtido com os dados de uma amostra de dimensão $n = 100$, simulada de uma distribuição binomial negativa, BiN(1,0.5).

do parâmetro de concentração c. Neste caso, o modelo Poisson é validado para valores pequenos de c, enquanto que, à medida que c aumenta, como esperado, o factor de Bayes fraccionário hierárquico aumenta e aproxima-se da unidade. No entanto, quando a amostra é simulada de uma das distribuições alternativas, binomial ou binomial negativa, o padrão de valores obtidos para o factor de Bayes fraccionário hierárquico é muito variado, dificultando a decisão de validação ou não do modelo em estudo.

Quando a amostra simulada é de uma distribuição alternativa mais afastada da distribuição de Poisson, como, por exemplo, as distribuições Bi(2,0.5), BiN(1,0.5), Bi(10,0.5) e BiN(5,0.5), todos os métodos rejeitam correctamente a adequabilidade do modelo Poisson, sendo que a estimativa do valor-p de discrepância e o factor pseudo-Bayes apresentam resultados mais significativos.

Por outro lado, quando a amostra simulada é de uma distribuição alternativa pró-

xima da distribuição de Poisson, como, por exemplo, as distribuições Bi(10,0.1), BiN(9,0.9), Bi(50,0.1) e BiN(45,0.9), nenhum método rejeita o modelo Poisson. Esta situação não é de estranhar uma vez que todos os modelos se adequam a dados discretos e as amostras simuladas, acima apresentadas, podem confundir-se facilmente com uma amostra da distribuição de Poisson.

À medida que a amostra simulada se afasta ligeiramente da distribuição de Poisson, como é o caso das amostras simuladas das distribuições Bi(4,0.25), BiN(2,2/3), Bi(20,0.25) e BiN(10,2/3), apenas o factor pseudo-Bayes e a estimativa do valor-p de discrepância conseguem dar alguma indicação de que o modelo Poisson não se adequa à amostra simulada.

Concluindo, de entre todos os métodos apresentados, a nova proposta, ou seja, o factor pseudo-Bayes, foi aquele que obteve os melhores resultados. O estudo da adequação de modelos para dados discretos, através de um estudo de simulação é, em nosso entender, útil e importante para tirar conclusões mais objectivas sobre o desempenho de cada um dos métodos e estudar a sua potência.

3.3.3 Dados contínuos

Seja $X = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória, referente a dados contínuos, constituída por n observações independentes e identicamente distribuídas.

Berger e Guglielmi (2001) apresentam um teste de ajustamento bayesiano não paramétrico, onde testam uma distribuição paramétrica (modelo paramétrico em estudo) contra uma distribuição não paramétrica (modelo não paramétrico). Por forma a incorporar o modelo paramétrico no modelo não paramétrico, os autores utilizam o modelo não paramétrico definido por uma mistura finita de árvores de Pólya (ver secção 2.2.3) centrada no modelo paramétrico. Seguidamente, utilizam o factor de Bayes simples como medida de comparação dos dois modelos (paramétrico *vs* não paramétrico). Apresenta-se, seguidamente, a construção deste teste de ajustamento.

Considere-se que o modelo bayesiano paramétrico é definido por

$$X_i|\theta \overset{\text{iid}}{\sim} f(x_i|\theta), \text{ para } i = 1, 2, \ldots, n$$
$$\theta \sim h(\theta)$$

e o modelo bayesiano não paramétrico é definido por

$$X_1, X_2, \ldots, X_n|G \overset{\text{iid}}{\sim} G$$
$$G|\Pi, \mathcal{A}_\theta \sim \text{MPT}_M(\Pi, \mathcal{A}_\theta),$$
$$\theta \sim h(\theta)$$

onde as partições $\Pi = \{B_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$ são fixas e os parâmetros da distribuição beta, $\mathcal{A}_\theta = \{\alpha_{\varepsilon_{1:m}}, m = 1, 2, \ldots, M\}$, dependem da distribuição e correspondente vector de parâmetros do modelo paramétrico, tal que se verifique a igualdade $E[G(B_{\varepsilon_{1:m}})|\theta] = F_\theta(B_{\varepsilon_{1:m}}) = \Pr(X_i \in B_{\varepsilon_{1:m}}|\theta)$.

Construção das partições

Por exemplo, se $X_i \in \mathbb{R}$ tem-se no primeiro nível da árvore $B_0 = (-\infty, F^{-1}_{\hat{\theta}}(0.5)]$ e $B_1 = (F^{-1}_{\hat{\theta}}(0.5), +\infty)$. Generalizando, no m-ésimo nível tem-se

$$B_{\varepsilon_{1:m}} = \left\{ \left(F^{-1}_{\hat{\theta}} \left(\frac{k-1}{2^m} \right), F^{-1}_{\hat{\theta}} \left(\frac{k}{2^m} \right) \right] \right\}, \tag{3.12}$$

para $m = 1, 2, \ldots, M$ e $k = 1, 2, \ldots, 2^m$, onde $F^{-1}_{\hat{\theta}}$ é a função quantil de X_i, substituindo o vector de parâmetros θ por estimativas de máxima verossimilhança.

Parâmetros da distribuição beta

Para $\varepsilon_{1:m-1} = \varepsilon_1 \varepsilon_2 \cdots \varepsilon_{m-1}$, os parâmetros da distribuição beta são definidos por

$$\alpha_{\varepsilon_{1:m-1}0}(\theta) = c_m \left(\frac{F_\theta(B_{\varepsilon_{1:m-1}0})}{F_\theta(B_{\varepsilon_{1:m-1}1})} \right)^{1/2} \tag{3.13}$$

e

$$\alpha_{\varepsilon_{1:m-1}1}(\theta) = c_m \left(\frac{F_\theta(B_{\varepsilon_{1:m-1}1})}{F_\theta(B_{\varepsilon_{1:m-1}0})} \right)^{1/2}, \tag{3.14}$$

onde

$$c_m \propto \eta^{-1}\rho(m), \;\; \eta > 0.$$

Desta forma, como $G(B_0|\theta) = Y_0 \sim \text{Be}(\alpha_0(\theta), \alpha_1(\theta))$, então

$$E[G(B_0)|\theta] = E[Y_0|\theta] = \frac{\alpha_0(\theta)}{\alpha_0(\theta) + \alpha_1(\theta)} = F_\theta(B_0),$$

e, para qualquer $B_{\varepsilon_{1:m}} \in \Pi$, $E[G(B_{\varepsilon_{1:m}})|\theta] = F_\theta(B_{\varepsilon_{1:m}})$, ou seja, o modelo não paramétrico é centrado no modelo paramétrico.

A função $\rho(m)$ é definida por forma a que a distribuição mistura finita de árvores de Pólya se adapte a distribuições amostrais contínuas, por exemplo, considerando $\rho(m) = m^2$, m^3, 2^m, 4^m, e 8^m. O parâmetro η (que coincide com c^{-1}, onde c é o parâmetro de concentração do processo de Dirichlet, ver secção 2.2.1), controla a variância da distribuição mistura finita de árvores de Pólya em torno da sua média (a distribuição paramétrica) e é de difícil especificação. Mais detalhes sobre este parâmetro η são dados mais à frente.

O teste de ajustamento bayesiano de Berger e Guglielmi (2001) pode ser definido por

$$H_0 : X \sim f(x|\theta) \;\; vs \;\; H_1 : X \sim G|\Pi, \mathcal{A}_\theta \;\; \theta \in \Theta$$

onde θ segue uma distribuição *a priori*, $h(\theta)$, usualmente não informativa. O factor de Bayes simples a favor do modelo paramétrico (H_0) e contra o modelo não paramétrico (H_1) é dado por

$$\text{BF}_{01}(x) = \frac{p_0(x)}{p_1(x)}. \tag{3.15}$$

A distribuição preditiva *a priori* no numerador (sob H_0) de (3.15), é dada por

$$p_0(x) = \int_\Theta f(x|\theta)h(\theta)d\theta,$$

onde $x = (x_1, x_2, \ldots, x_n)$ e $f(x|\theta) = \prod_{i=1}^n f(x_i|\theta)$.

A distribuição preditiva *a priori* no denominador (sob H_1) de (3.15), e segundo Lavine (1992), é definida por

$$p_1(x) = \int_\Theta p(x|\theta)h(\theta)d\theta,$$

onde

$$p(x|\theta) = f(x|\theta)\psi(\theta), \tag{3.16}$$

com

$$\psi(\theta) = \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha'_{\varepsilon_{1:m}(x_j)}(\theta)\left(\alpha_{\varepsilon_{1:m-1}0(x_j)}(\theta) + \alpha_{\varepsilon_{1:m-1}1(x_j)}(\theta)\right)}{\alpha_{\varepsilon_{1:m}(x_j)}(\theta)\left(\alpha'_{\varepsilon_{1:m-1}0(x_j)}(\theta) + \alpha'_{\varepsilon_{1:m-1}1(x_j)}(\theta)\right)}, \tag{3.17}$$

onde $\varepsilon_{1:m}(x_j)$ é o índice $\varepsilon_1\varepsilon_2\cdots\varepsilon_m$ que identifica o subconjunto da partição $B_{\varepsilon_1\cdots\varepsilon_m}$, para cada nível m, que contém x_j, e $\alpha'_{\varepsilon_{1:m}(x_j)}(\theta)$ é igual a $\alpha_{\varepsilon_{1:m}(x_j)}(\theta)$ mais o número de observações entre $\{x_1, \ldots, x_{j-1}\}$ que pertencem a $B_{\varepsilon_1\cdots\varepsilon_m}(x_j)$. Para cada x_j, o limite superior $m^*(x_j)$ no produto em (3.17) representa o menor nível m tal que nenhum x_i, $i < j$, pertence a $B_{\varepsilon_1\cdots\varepsilon_m}(x_j)$.

O cálculo do factor de Bayes pode ser simplificado porque por, (3.16), o factor de Bayes, definido em (3.15), pode ser escrito como

$$\mathrm{BF}_{01}(x) = \left[\int_\Theta \psi(\theta)h(\theta|x)d\theta\right]^{-1} = \left\{E[\psi(\theta)|x]\right\}^{-1},$$

onde $h(\theta|x) = f(x|\theta)h(\theta)/p_0(x)$, isto é, pode ser escrito como o inverso de uma média *a posteriori*, sob H_0.

Caso se possa simular uma amostra aleatória $\theta_1, \theta_2, \ldots, \theta_L$ da densidade *a posteriori* $h(\theta|x)$, o método de Monte Carlo directo aproxima o factor de Bayes pelo inverso da média empírica

$$\widehat{\mathrm{BF}}_{01}(x) = \left[\frac{1}{L}\sum_{l=1}^{L}\psi(\theta_l)\right]^{-1}. \tag{3.18}$$

A precisão desta aproximação pode ser medida pelo erro padrão (estimado) de Monte Carlo, dado por

$$ep = \frac{1}{\sqrt{L(L-1)}}\left\{\sum_{l=1}^{L}\left[\psi(\theta_l) - \frac{1}{L}\sum_{l=1}^{L}\psi(\theta_l)\right]^2\right\}^{1/2}.$$

O método de Monte Carlo directo pode ser de convergência lenta, particularmente se a distribuição *a posteriori* tiver caudas leves. Para contornar este problema e aumentar

a velocidade de convergência, pode sempre que possível, utilizar-se o método de Monte Carlo com amostragem via função de importância.

Se existir uma função de importância $q_0(\theta)$, que tenha caudas mais pesadas, fácil de simular e que aproxima a distribuição *a posteriori* $h(\theta|x)$, então se se obtiver uma amostra $\theta_1, \theta_2, \ldots, \theta_L$ de $q_0(\theta)$, pode aplicar-se o método de Monte Carlo, e o valor aproximado para o factor de Bayes é dado por

$$\widehat{\mathrm{BF}}_{01}(x) = \left[\frac{1}{\sum_{l=1}^{L} w(\theta_l)} \sum_{l=1}^{L} w(\theta_l)\psi(\theta_l) \right]^{-1}, \tag{3.19}$$

onde $w(\theta_l) = f(x|\theta_l)h(\theta_l)/q_0(\theta_l)$, com um erro padrão de Monte Carlo estimado dado por

$$ep = \tfrac{1}{\sum_{l=1}^{L} w(\theta_l)} \times \times \left[\sum_{l=1}^{L} \left\{ \psi(\theta_l) - \frac{1}{\sum_{l=1}^{L} w(\theta_l)} \sum_{l=1}^{L} w(\theta_l)\psi(\theta_l) \right\}^2 w(\theta_l)^2 \right]^{1/2}.$$

As distribuições multivariadas adequadas para funcionar como função de importância são, usualmente, distribuições normais multivariadas ou distribuições Student multivariadas.

Berger e Guglielmi (2001) propõem analisar graficamente as estimativas do factor de Bayes como função do parâmetro η, uma vez que a variação dos valores de η determina quão concentrada a distribuição não paramétrica (mistura finita de árvores de Pólya) está da sua correspondente média (a distribuição paramétrica). Para valores de $\eta \to 0$, a distribuição não paramétrica está mais concentrada em torno da distribuição paramétrica e o factor de Bayes irá convergir para um. Para valores de $\eta \to \infty$, a distribuição não paramétrica está mais afastada da distribuição paramétrica e o factor de Bayes será muito grande. Entre estes dois extremos, o factor de Bayes, por vezes, aumenta com η, mas também pode, inicialmente, diminuir para depois aumentar. Por conseguinte, os autores optam por uma análise de robustez, calculando o factor de Bayes a favor do modelo paramétrico (H_0) e contra o modelo não paramétrico (H_1) para vários valores de η e, seguidamente, escolhem o valor mínimo obtido, $\min(\widehat{\mathrm{BF}}_{01}(x))$, como uma escolha

conservativa (Tokdar e Martin, 2011).

3.3.4 Exemplos de aplicação

Muitos dos procedimentos estatísticos, pressupõem que os dados são normalmente distribuídos. Nesta secção, exemplificam-se com base em dois conjuntos de dados, como se aplica o teste de ajustamento bayesiano não paramétrico de Berger e Guglielmi (2001), tendo como principal objectivo entender como se pode generalizar o referido teste para outras distribuições, para além da distribuição normal.

O primeiro conjunto de dados consiste numa amostra de 100 observações simuladas de uma distribuição normal, N(100,10). O segundo conjunto de dados, encontra-se em Andrews e Herzberg (1985, pag. 183), e consiste em n =100 tempos de vida até à ruptura de uma liga de Kevlar (depois de aplicar uma transformação logarítmica aos dados originais). Pretende-se testar, se a distribuição normal se adequa a cada um dos dois conjuntos de dados.

Representa-se graficamente, e para cada um dos dois conjuntos de dados separadamente, o histograma com sobreposição da densidade estimada e o gráfico dos quantis empíricos contra os quantis teóricos. Verifica-se que a forma simétrica da distribuição normal parece estar presente no histograma da Figura 3.3, como esperado, mas não no histograma da Figura 3.4, onde parece haver um enviesamento à esquerda. Verifica-se, também, que no gráfico dos quantis da Figura 3.3, os pontos se aproximam bem da recta de referência, havendo um grande desvio dos pontos em relação à mesma recta, no gráfico dos quantis da Figura 3.4.

Seguidamente, apresentam-se os passos dados por Berger e Guglielmi (2001) para realizar o teste de ajustamento bayesiano não paramétrico, para estudar a adequabilidade da distribuição normal, utilizando os dois conjuntos de dados.

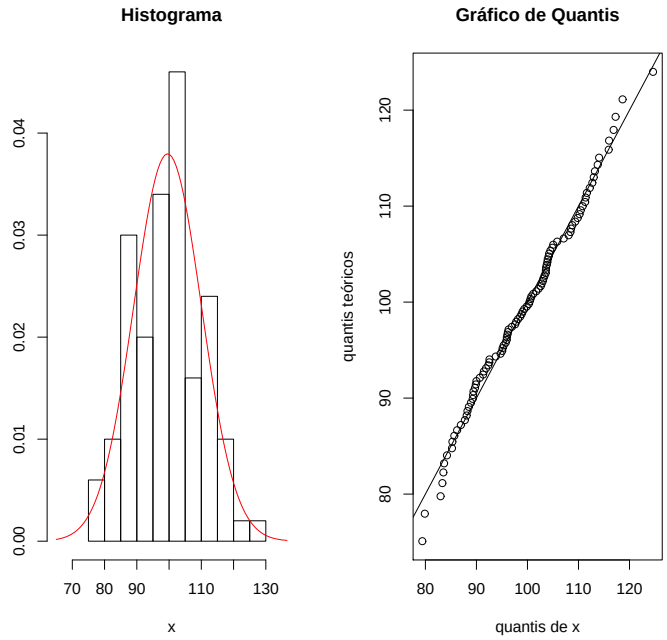

Figura 3.3: Histograma com sobreposição da densidade estimada (esquerda) e gráfico dos quantis empíricos contra os quantis teóricos (direita) das 100 observações simuladas de uma distribuição normal, N(100,10).

O modelo bayesiano paramétrico é dado por

$$X_i|\theta = (\mu, \sigma^2) \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \text{ para } i = 1, 2, \ldots, n,$$
$$\theta \sim h(\theta) \propto \tfrac{1}{\sigma}$$

e o modelo bayesiano não paramétrico é

$$X_1, X_2, \ldots, X_n|G \overset{\text{iid}}{\sim} G$$
$$G|\Pi, \mathcal{A}_\theta \sim \text{MPT}_M(\Pi, \mathcal{A}_\theta),$$
$$\theta \sim h(\theta) \propto \tfrac{1}{\sigma}$$

onde $\text{MPT}_M(\Pi, \mathcal{A}_\theta)$ define uma distribuição *a priori* mistura finita de árvores de Pólya, com parâmetros $(\Pi, \mathcal{A}_\theta)$ e M níveis pré-especificados, centrada no modelo paramétrico, $N(\mu, \sigma^2)$, com informação *a priori* não informativa para o correspondente vector de parâmetros $\theta = (\mu, \sigma^2)$.

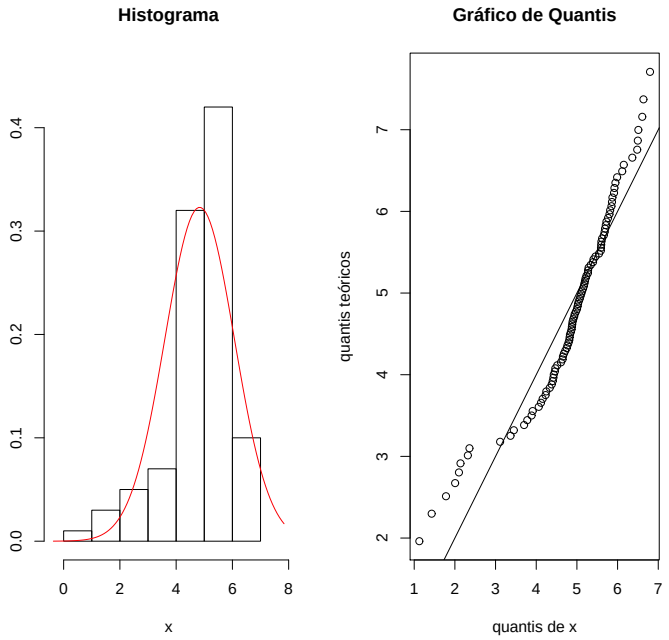

Figura 3.4: Histograma com sobreposição da densidade estimada (esquerda) e gráfico dos quantis empíricos contra os quantis teóricos (direita) dos 100 tempos de vida até à ruptura de uma liga de Kevlar.

Para o cálculo do factor de Bayes, Berger e Guglielmi (2001) optam pela aproximação via método de Monte Carlo com amostragem via função de importância, cuja a expressão está definida em (3.19).

A função de importância, $q_0(\mu, \sigma^2)$, utilizada pelos referidos autores é

$$q_0(\mu, \sigma^2) = \frac{n}{\sqrt{2}\pi\kappa\hat{\sigma}\sigma^2}\left[1 + \frac{n}{2\kappa\hat{\sigma}^2}(\mu - \hat{\mu})^2 + \frac{n}{4\kappa}(\log\frac{\sigma^2}{\hat{\sigma}^2})^2\right]^{-3}, \tag{3.20}$$

onde $\hat{\mu}$ e $\hat{\sigma}^2$ são as correspondentes estimativas de máxima verosimilhança.

A função de importância (3.20) é obtida considerando que o vector $(\mu, \ln(\sigma^2))$ é caracterizado por uma função densidade de uma distribuição t-Student bivariada com 4 graus de liberdade, com vector médio dado por $(\bar{x}, \log(s^2))$ e matriz de variâncias e covariâncias dada por $\Sigma_{(\mu,\log(\sigma^2))} = \kappa\hat{I}^{-1}$, onde κ é uma constante positiva e $\hat{I}$ corresponde

à matriz de informação de Fisher, isto é,

$$\hat{I} = \begin{bmatrix} \frac{n}{s^2} & 0 \\ 0 & \frac{n}{2} \end{bmatrix}.$$

Aplicando a transformação inversa, obtém-se $q_0(\mu, \sigma^2)$. A constante κ controla a velocidade de convergência da função de importância. De acordo com estudos de simulação efectuados por Berger e Guglielmi (2001), valores de κ entre 10 e 100 são bastante aceitáveis. Daí os autores optarem por tomar $\kappa = 50$ para o cálculo do valor estimado do factor de Bayes.

Partindo de um vector de dados observados $x = (x_1, x_2, \ldots, x_n)$, o cálculo da estimativa do factor de Bayes, a favor do modelo paramétrico e contra o modelo não paramétrico, pode ser sumariado através do seguinte algoritmo:

Algoritmo 3

1. Define-se o número M de níveis da árvore (por exemplo, utilizando a sugestão de Hanson e Johnson (2002), $M \simeq \log_2(n)$);

2. Calculam-se as partições binárias do espaço amostral $\mathbb{R}$ para os M níveis, dadas por
$$B_{\varepsilon_{1:m}} = \left\{ \left(F_{\hat{\theta}}^{-1}\left(\frac{k-1}{2^m}\right), F_{\hat{\theta}}^{-1}\left(\frac{k}{2^m}\right) \right] \right\},$$
para $m = 1, 2, \ldots, M$ e $k = 1, 2, \ldots, 2^m$, onde $F_{\hat{\theta}}^{-1}(\cdot)$ são os quantis da distribuição normal tomando para valores dos parâmetros as respectivas estimativas de máxima verosimilhança, $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, s^2)$, com $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ e $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$;

3. Define-se a(s) expressão(ões) de $\rho(m)$ e o intervalo de valores para η;

4. Para cada valor de η,

   (a) Para $l = 1, 2, \ldots, L$

      i. Gera-se um vector aleatório θ_l da função de importância $q_0(\theta)$, $\theta = (\mu, \sigma^2)$;

ii. Calculam-se os valores para os parâmetros da distribuição beta:

$$\alpha_{\varepsilon_{1:m-1}0}(\theta_l) = c_m \left(\frac{F_{\theta_l}(B_{\varepsilon_{1:m-1}0})}{F_{\theta_l}(B_{\varepsilon_{1:m-1}1})} \right)^{1/2}$$

e

$$\alpha_{\varepsilon_{1:m-1}1}(\theta_l) = c_m \left(\frac{F_{\theta_l}(B_{\varepsilon_{1:m-1}1})}{F_{\theta_l}(B_{\varepsilon_{1:m-1}0})} \right)^{1/2},$$

onde

$$c_m \propto \eta^{-1}\rho(m);$$

(b) Calcula-se uma estimativa do factor de Bayes, utilizando a expressão definida em (3.19);

5. Determina-se o valor mínimo das estimativas do factor de Bayes calculadas.

Utilizam-se $L = 10000$ iterações uma vez que permite obter um erro padrão estimado de Monte Carlo pequeno e consideram-se árvores de Pólya com $M = 8$ níveis. Quanto à definição das expressões de $\rho(m)$ e aos valores a atribuir ao parâmetro η, os referidos autores optam por variar η entre os inteiros de 1 a 100 e utilizar diferentes expressões para $\rho(m)$, tais como, 2^m, 4^m, 8^m, etc..

Teoricamente, e se a distribuição paramétrica se adequa ao conjunto de dados, então, para valores de $\eta \to 0$, a distribuição não paramétrica está mais concentrada em torno da distribuição paramétrica e, consequentemente, o factor de Bayes estará próximo de um. Daí, seguindo a sugestão de Tokdar e Martin (2011), também se apresentam as estimativas do factor de Bayes para valores de η inferiores, iguais e superiores a um, num total de apenas 13 estimativas (e não as 100 estimativas anteriores). Para isso, define-se $\eta = 2^s$, com s a tomar todos os valores inteiros pertencentes ao intervalo $[-6, 6]$.

Rejeita-se a hipótese da amostra observada seguir uma distribuição normal se a estimativa final do factor de Bayes, a favor do modelo paramétrico e contra o modelo não paramétrico, $\min(\widehat{\mathrm{BF}}_{01}(x))$, assumir valores pequenos (inferiores a um) ou se a

estimativa final do logaritmo de base 10 do factor de Bayes, $\min(\log_{10}(\widehat{\mathrm{BF}}_{01}))$, for negativa.

Nas Figuras 3.5 a 3.8 encontram-se as representações gráficas das 100 e das 13 estimativas do logaritmo do factor de Bayes, respectivamente, para a amostra simulada de uma distribuição normal e para a amostra dos tempos de vida, e para diferentes expressões de c_m: $\frac{10}{\eta}2^m$, $\frac{1}{\eta}4^m$, $\frac{1}{\eta}8^m$ e $\frac{10}{\eta}m^2$. Nas Tabelas 3.4 a 3.7 apresentam-se as correspondentes estimativas finais do factor de Bayes.

De acordo com os resultados obtidos, conclui-se que o teste de ajustamento bayesiano não paramétrico não rejeita a hipótese de normalidade para o conjunto de dados simulados e rejeita a hipótese de normalidade para a amostra dos tempos de vida, como se esperaria.

Relativamente à utilização das duas propostas para definir os diferentes valores para η, verifica-se que, quando o modelo é rejeitado, não há diferenças significativas nos valores obtidos. No entanto, quando a distribuição paramétrica se adequa ao conjunto de dados em estudo, valores de η inferiores a um e próximos de zero confirmam a conclusão teórica que já se mencionou, isto é, o factor de Bayes está próximo de um. Esta conclusão é importante na medida que pode ser relevante aquando, por exemplo, da definição de um valor de corte (*threshold*) para o factor de Bayes num estudo de simulação, como veremos mais à frente.

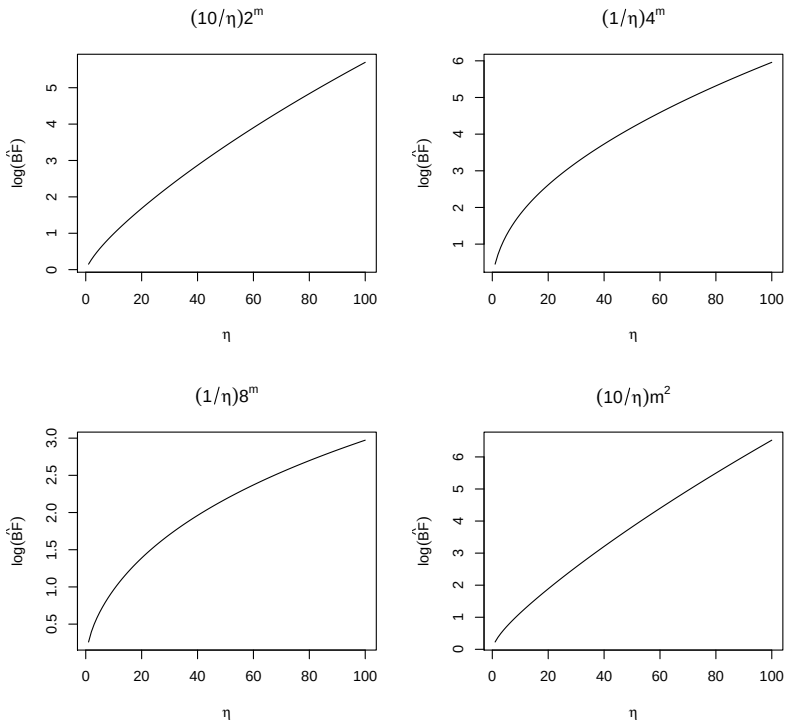

Figura 3.5: Representação gráfica das 100 estimativas do logaritmo do factor de Bayes para a amostra simulada de uma distribuição normal, N(100,10), e para 4 diferentes expressões de c_m.

Tabela 3.4: Valor mínimo das 100 estimativas do factor de Bayes e do seu logaritmo para a amostra simulada de uma distribuição normal, N(100,10), e para 4 diferentes expressões de c_m.

c_m	$\min(\log(\widehat{\mathrm{BF}}_{01}))$	$\min(\widehat{\mathrm{BF}}_{01})$
$\frac{10}{\eta}2^m$	0.1508	1.4153
$\frac{1}{\eta}4^m$	0.4532	2.8390
$\frac{1}{\eta}8^m$	0.2595	1.8118
$\frac{10}{\eta}m^2$	0.2279	1.6904

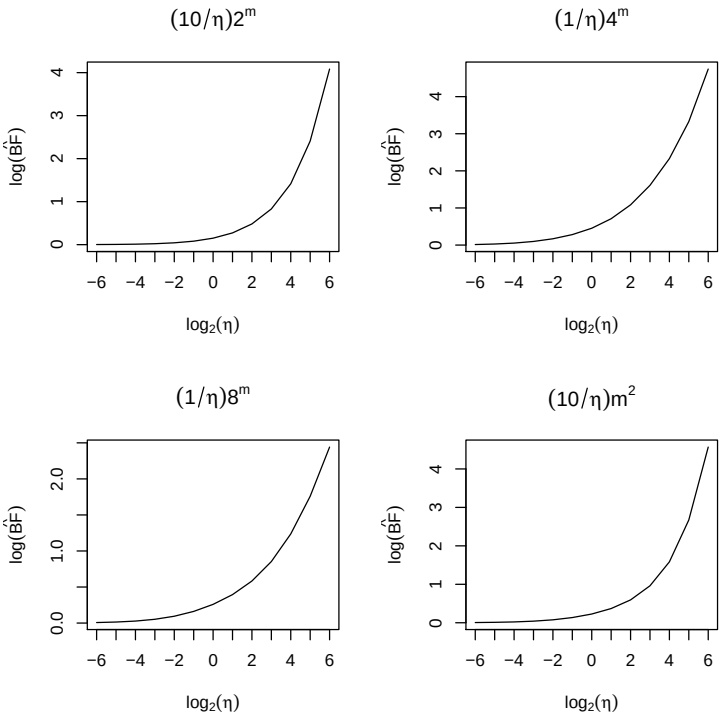

Figura 3.6: Representação gráfica das 13 estimativas do logaritmo do factor de Bayes para a amostra simulada de uma distribuição normal, N(100,10), e para 4 diferentes expressões de c_m.

Tabela 3.5: Valor mínimo das 13 estimativas do factor de Bayes e do seu logaritmo para a amostra simulada de uma distribuição normal, N(100,10), e para 4 diferentes expressões de c_m.

c_m	$\min(\log(\widehat{\mathrm{BF}}_{01}))$	$\min(\widehat{\mathrm{BF}}_{01})$
$\frac{10}{\eta}2^m$	0.0028	1.0065
$\frac{1}{\eta}4^m$	0.0144	1.0339
$\frac{1}{\eta}8^m$	0.0073	1.0169
$\frac{10}{\eta}m^2$	0.0056	1.0130

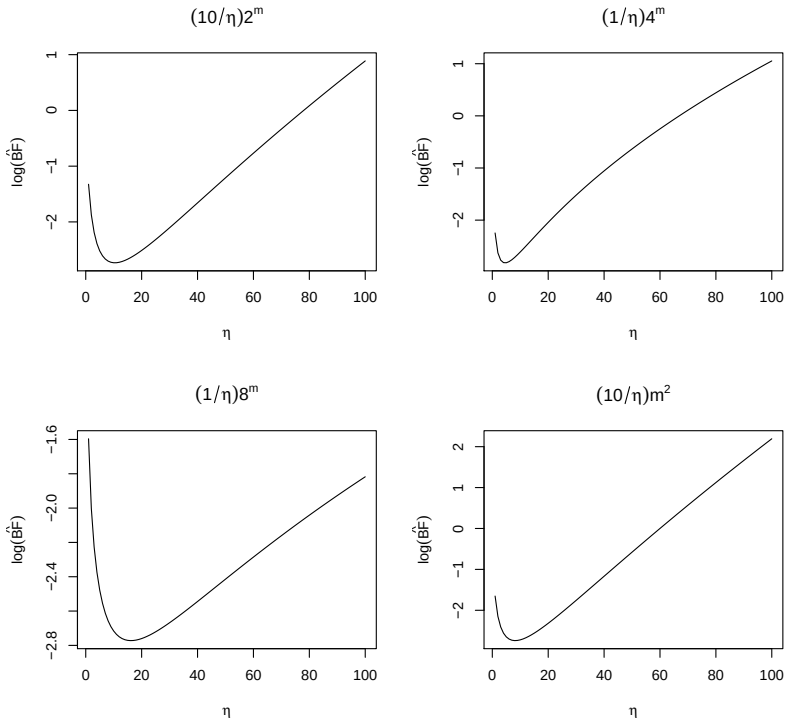

Figura 3.7: Representação gráfica das 100 estimativas do logaritmo do factor de Bayes para a amostra dos tempos de vida e para 4 diferentes expressões de c_m.

Tabela 3.6: Valor mínimo das 100 estimativas do factor de Bayes e do seu logaritmo para a amostra dos tempos de vida e para 4 diferentes expressões de c_m.

c_m	$\min(\log(\widehat{\mathrm{BF}}_{01}))$	$\min(\widehat{\mathrm{BF}}_{01})$
$\frac{10}{\eta}2^m$	-2.7327	0.0018
$\frac{1}{\eta}4^m$	-2.8174	0.0015
$\frac{1}{\eta}8^m$	-2.7722	0.0017
$\frac{10}{\eta}m^2$	-2.7413	0.0018

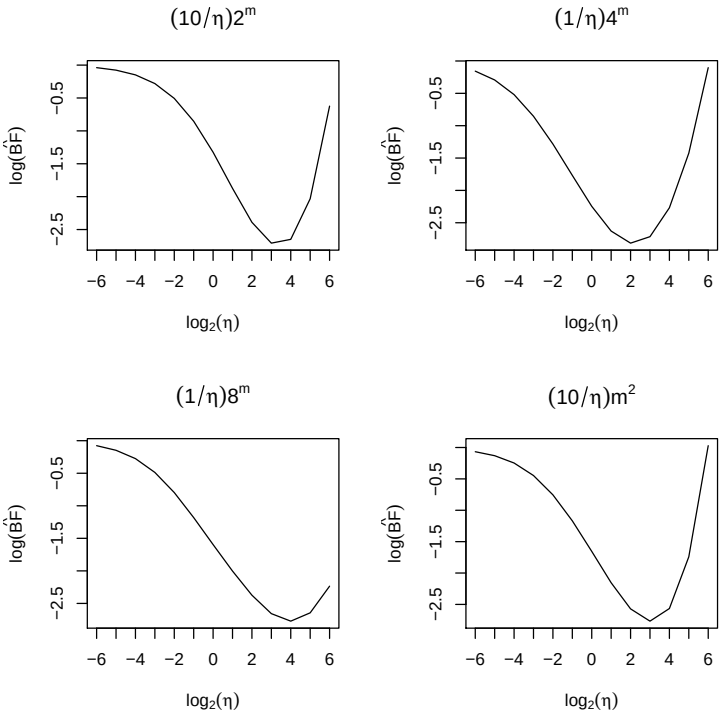

Figura 3.8: Representação gráfica das 13 estimativas do logaritmo do factor de Bayes para a amostra dos tempos de vida e para 4 diferentes expressões de c_m.

Tabela 3.7: Valor mínimo das 13 estimativas do factor de Bayes e do seu logaritmo para a amostra dos tempos de vida e para 4 diferentes expressões de c_m.

c_m	$\min(\log(\widehat{\mathrm{BF}}_{01}))$	$\min(\widehat{\mathrm{BF}}_{01})$
$\frac{10}{\eta}2^m$	-2.7067	0.0019
$\frac{1}{\eta}4^m$	-2.8147	0.0015
$\frac{1}{\eta}8^m$	-2.7722	0.0017
$\frac{10}{\eta}m^2$	-2.7666	0.0017

Capítulo 4

O modelo exponencial

A distribuição exponencial é uma das mais simples e importantes distribuições e é utilizada na modelação de dados que representam o tempo até à ocorrência de um determinado acontecimento de interesse, muitas vezes denominado por tempo de vida ou tempo até à falha. Este tipo de dados é frequente em análise de sobrevivência e em análise de fiabilidade, entre outras áreas. O estudo da adequabilidade da distribuição exponencial a um conjunto de dados observado é fundamental para que as inferências depois realizadas sejam válidas.

Uma variável aleatória, não negativa e contínua, X, que representa o tempo até à ocorrência de um determinado acontecimento de interesse, tem distribuição exponencial, $X \sim \text{Exp}(\lambda)$, com parâmetro $\lambda > 0$, se a sua função densidade é dada por

$$f(x|\lambda) = \lambda \exp(-\lambda x), \quad x \geq 0$$

onde o parâmetro λ é designado de taxa de falha da distribuição.

A abordagem clássica para o estudo da adequabilidade da distribuição exponencial tem sido um tema bastante estudado. Veja-se, por exemplo, os trabalhos de Baringhaus e Henze (1991, 2000), Choi et al. (2004), Henze e Meintanis (2005), Grané e Fortiana (2011), assim como outras referências importantes nesses artigos. Os testes de ajustamento clássicos usuais utilizam estatísticas baseadas, por exemplo, na função de distri-

buição empírica, como é o caso do teste de Kolmogorov-Smirnov, o teste de Cramér-von Mises e o teste de Anderson-Darling. Outras estatísticas têm sido desenvolvidas e, paralelamente, têm sido realizados estudos de simulação para avaliar o comportamento de cada um dos testes quanto à sua taxa de erro tipo I e à sua potência.

Relativamente à abordagem bayesiana para o estudo da adequabilidade da distribuição exponencial a um conjunto de dados, segundo a filosofia dos testes de ajustamento, não se conhecem trabalhos nesta área, assim como não se conhecem estudos que permitam fazer comparações relevantes entre os testes clássicos e os testes bayesianos. Propõe-se, neste trabalho, o desenvolvimento de alguns testes de ajustamento e a sua comparação, através de um estudo de simulação Monte Carlo, com alguns testes clássicos, estes últimos seleccionados na literatura estatística e considerados como os mais potentes.

4.1 Testes clássicos

Um grande número de testes clássicos para o estudo da adequabilidade da distribuição exponencial têm sido propostos na literatura. Estes testes baseiam-se em diferentes características da distribuição exponencial e podem ser classificados em várias categorias. Henze e Meintanis (2005) identificam oito categorias de testes e comparam vinte e uma estatísticas de teste para o estudo da adequabilidade da distribuição exponencial contra dezoito distribuições alternativas, através de um estudo de simulação Monte Carlo. O estudo exaustivo dos referidos autores permite concluir que não existe uma estatística de teste melhor, em termos de potência, que todas as outras estudadas. No entanto, o estudo dá indicações que a estatística de Cox e Oakes (1984), CO_n, a estatística de Epps e Pulley (1986), EP_n, a estatística de teste clássica de Cramér-von Mises modificada, $\overline{\mathrm{CM}}_n$, de Baringhaus e Henze (2000), baseada numa caracterização da função de sobrevivência média residual, a estatística de teste clássica, $\mathrm{BH}_{n,a}$, de Baringhaus e Henze (1991), baseada na transformada de Laplace empírica e a nova estatística de

teste apresentada por Henze e Meintanis (2005), $T_{n,a}$, baseada na função característica empírica, estão entre as mais potentes e simples de calcular. As duas últimas estatísticas são dependentes de uma constante arbitrária a e a sua potência é drasticamente afectada pelo seu valor. Seguindo a sugestão dos referidos autores, utiliza-se $a = 1$, $a = 1.5$ e $a = 2.5$ para $\mathrm{BH}_{n,a}$ e $a = 1.5$ e $a = 2.5$ para $T_{n,a}$, esta última apenas aplicada a amostras de pequena dimensão ($n = 25$ e $n = 50$). A estatística de teste de Anderson e Darling (1954), AD_n, não foi incluída no estudo de Henze e Meintanis (2005) mas, como é uma estatística muito utilizada na literatura, inclui-se também no presente trabalho.

Definem-se, seguidamente, cada uma das estatísticas de teste clássicas, apresentando apenas as expressões práticas utilizadas para o estudo da adequabilidade da distribuição exponencial, usualmente associadas às hipóteses estatísticas, $H_0 : X \sim \mathrm{Exp}(\lambda)$ *vs* $H_1 : X \nsim \mathrm{Exp}(\lambda)$. Todos os pormenores teóricos relativos às estatísticas de teste apresentadas podem ser estudados, com detalhe, nas referências indicadas.

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória, constituída por n observações independentes e identicamente distribuídas a X, e seja $Y_i = X_i/\bar{X}$ com $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$.

1. A estatística de teste de Cox e Oakes (1984) é dada por

$$\mathrm{CO}_n = n + \sum_{i=1}^{n} (1 - Y_i)\log(Y_i).$$

2. A estatística de teste de Epps e Pulley (1986) é definida por

$$\mathrm{EP}_n = (48n)^{1/2} + \left[\frac{1}{n}\sum_{i=1}^{n}\exp(-Y_i) - \frac{1}{2}\right].$$

3. A estatística de teste de Cramér-von Mises modificada de Baringhaus e Henze (2000) é calculada como

$$\overline{\mathrm{CM}}_n = \frac{1}{n}\sum_{i,k=1}^{n}\Big(2 - 3e^{-\min(Y_i,Y_k)} \\ -2\min(Y_i, Y_k)(e^{-Y_i} + e^{-Y_k}) + 2e^{-\max(Y_i,Y_k)}\Big).$$

4. A estatística de teste de Baringhaus e Henze (1991) é definida por

$$\mathrm{BH}_{n,a} = \frac{1}{n}\sum_{i,k=1}^{n}\left[\frac{(1-Y_i)(1-Y_k)}{Y_i+Y_k+a} - \frac{Y_i+Y_k}{(Y_i+Y_k+a)^2} + \frac{2Y_iY_k}{(Y_i+Y_k+a)^2} + \frac{2Y_iY_k}{(Y_i+Y_k+a)^3}\right].$$

5. A nova estatística de teste de Henze e Meintanis (2005) é dada por

$$\begin{aligned}\mathrm{T}_{n,a} = {} & \frac{a}{n}\sum_{i,k=1}^{n}\left[\frac{1}{a^2+(Y_i-Y_k)^2} + \frac{1}{a^2+(Y_i+Y_k)^2}\right] \\ & -\frac{2a}{n^2}\sum_{i,k=1}^{n}\sum_{l=1}^{n}\left[\frac{1}{a^2+(Y_i-Y_k-Y_l)^2} + \frac{1}{a^2+(Y_i+Y_k+Y_l)^2}\right] \\ & +\frac{a}{n^3}\sum_{i,k=1}^{n}\sum_{l,m=1}^{n}\left[\frac{1}{a^2+(Y_i-Y_k-(Y_l-Y_m))^2} \right. \\ & \left. +\frac{1}{a^2+(Y_i+Y_k+(Y_l-Y_m))^2}\right].\end{aligned}$$

6. A estatística de teste de Anderson e Darling (1954) é dada por

$$\begin{aligned}\mathrm{AD}_n = {} & -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1) \\ & \times\left[\log(W_{(i)}) + \log(1-W_{(n-i+1)})\right],\end{aligned}$$

onde $W_{(i)} = 1-\exp(-Y_{(i)})$, $1 \leq i \leq n$, e $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ definem as estatísticas ordinais de $(Y_1, Y_2, \ldots, Y_n)$.

Para as duas primeiras estatísticas de teste, os respectivos autores provam que, sob a hipótese nula, $X \sim \mathrm{Exp}(\lambda)$, EP_n e $\mathrm{CO}_n^* = \left(\frac{6}{n}\right)^{1/2}\left(\frac{\mathrm{CO}_n}{\pi}\right)$ seguem uma distribuição assintótica normal reduzida e, portanto, é rejeitada a hipótese nula para grandes valores de $|\mathrm{EP}_n|$ e de $|\mathrm{CO}_n^*|$. Para as últimas quatro estatísticas de teste, as suas distribuições, sob a hipótese nula, não estão definidas analiticamente, sendo necessário recorrer à simulação Monte Carlo para obter os respectivos valores críticos. Rejeita-se a hipótese de a amostra observada seguir uma distribuição exponencial se o valor observado de cada uma das quatro estatísticas de teste exceder o seu valor crítico empírico.

Para obter os valores críticos empíricos de cada uma das quatro estatísticas clássicas, $\overline{\mathrm{CM}}_n$, AD_n, $\mathrm{BH}_{n,a}$ e $\mathrm{T}_{n,a}$, 100 000 amostras aleatórias de dimensão n foram simuladas da

distribuição exponencial padrão, Exp(1), e os respectivos valores das quatro estatísticas de teste calculados. As dimensões amostrais consideradas são $n = 25$, $n = 50$ e $n = 100$, e os níveis de significância escolhidos são $\alpha = 0.1, 0.05$ e 0.025 para as estatísticas $\overline{\mathrm{CM}}_n$, AD_n e $\mathrm{BH}_{n,a}$. Relativamente à estatística de teste $\mathrm{T}_{n,a}$, calcula-se os valores críticos empíricos apenas para as dimensões $n = 25$ e $n = 50$, assim como, apenas para os níveis de significância $\alpha = 0.1$ e 0.05. Os valores críticos empíricos para cada uma das estatísticas de teste, apresentados nas Tabelas 4.1, 4.2 e 4.3, foram calculados através dos quantis $(1 - \alpha)100\%$ da correspondente distribuição empírica. Chama-se a atenção para o facto de que, relativamente às estatísticas $\mathrm{BH}_{n,a}$ e $\mathrm{T}_{n,a}$, os valores críticos empíricos obtidos estarem muito próximos dos definidos pelos respectivos autores.

Tabela 4.1: Valores críticos empíricos das estatísticas de teste $\overline{\mathrm{CM}}_n$ e AD_n.

	$\overline{\mathrm{CM}}_n$			AD_n		
α	0.1	0.05	0.025	0.1	0.05	0.025
$n = 25$	0.341	0.451	0.564	1.042	1.292	1.562
$n = 50$	0.344	0.455	0.567	1.053	1.312	1.573
$n = 100$	0.348	0.464	0.583	1.058	1.325	1.595

Tabela 4.2: Valores críticos empíricos das estatísticas de teste $\mathrm{T}_{n,a}$, para $a = 1.5$ e $a = 2.5$.

	$\mathrm{T}_{n,a=1.5}$		$\mathrm{T}_{n,a=2.5}$	
α	0.1	0.05	0.1	0.05
$n = 25$	0.275	0.411	0.077	0.109
$n = 50$	0.256	0.359	0.075	0.104

Tabela 4.3: Valores críticos empíricos da estatística de teste $\mathrm{BH}_{n,a}$, para $a = 1$, $a = 1.5$ e $a = 2.5$.

	$\mathrm{BH}_{n,a=1}$			$\mathrm{BH}_{n,a=1.5}$			$\mathrm{BH}_{n,a=2.5}$		
α	0.1	0.05	0.025	0.1	0.05	0.025	0.1	0.05	0.025
$n = 25$	0.219	0.304	0.385	0.140	0.194	0.251	0.071	0.098	0.127
$n = 50$	0.220	0.305	0.396	0.142	0.199	0.258	0.073	0.102	0.131
$n = 100$	0.221	0.311	0.401	0.143	0.199	0.260	0.073	0.104	0.135

4.2 Testes bayesianos

No Capítulo 3 apresentaram-se, entre outros, dois testes bayesianos, nomeadamente o teste do qui-quadrado bayesiano de Johnson (2004) e o teste de ajustamento bayesiano não paramétrico de Berger e Guglielmi (2001). A teoria destes dois testes vai, agora, ser adaptada ao estudo da adequabilidade da distribuição exponencial a um conjunto de dados observados. Seguidamente, apresentam-se, resumidamente, as alterações a efectuar por forma a realizar cada um dos testes.

O teste do qui-quadrado bayesiano de Johnson (2004), apresentado na secção 3.2, aplicado ao estudo da adequabilidade da distribuição exponencial, vai ser caracterizado pela seguinte estatística de teste

$$Q_n^B(\check{\lambda}) = \sum_{j=1}^{k} \frac{(m_j(\check{\lambda}) - np_j)^2}{np_j} \overset{\circ}{\sim} \chi^2(k-1), \tag{4.1}$$

onde k é o número de sub-intervalos (classes) equiprováveis ($p_j = 1/k$), obtido utilizando a regra $k \cong n^{0.4}$, com n a representar a dimensão da amostra observada. $\check{\lambda}$ é um valor simulado da distribuição *a posteriori*, utilizando uma distribuição *a priori* não informativa da família conjugada natural, $\text{Ga}(a, b)$, com $a, b \to 0$, e $m_j(\check{\lambda})$ é o número de observações que caem na j-ésima classe, isto é, o número de observações que satisfazem $F(x_i|\check{\lambda}) \in (a_{j-1}, a_j]$, para $i = 1, 2, \ldots, n$.

Para verificar se a distribuição assintótica da estatística de teste definida em (4.1) se mantém como uma distribuição qui-quadrado com $(k-1)$ graus de liberdade, simulou-se 10000 amostras de dimensão $n = 100$ ($k = 6$) da distribuição exponencial padrão e calcula-se, para cada uma das amostras, a referida estatística de teste. Na Figura 4.1, do lado esquerdo, representa-se graficamente a densidade estimada dos 10000 valores obtidos de $Q_n^B(\check{\lambda})$, com sobreposição da função densidade de probabilidade da distribuição qui-quadrado com $(k-1=5)$ graus de liberdade. Do lado direito da Figura 4.1 representa-se graficamente os quantis empíricos contra os quantis teóricos.

Verifica-se que a densidade estimada está muito próxima da densidade teórica e

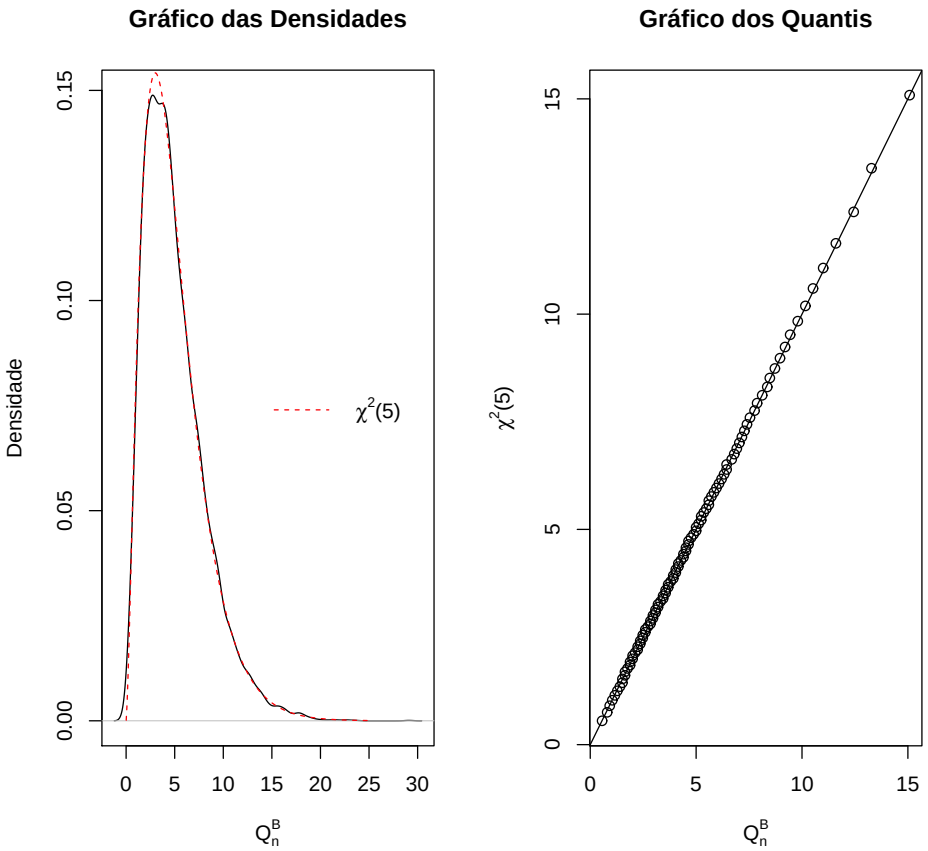

Figura 4.1: Densidade estimada de $Q_n^B(\check{\lambda})$ com sobreposição da densidade da distribuição qui-quadrado, $\chi^2(5)$, (à esquerda) e gráfico dos quantis empíricos contra os quantis teóricos (à direita).

os pontos representados no gráfico dos quantis estão muito próximos da bissetriz do primeiro quadrante. Conclui-se, assim, que a distribuição assintótica da estatística de teste do qui-quadrado bayesiano para o estudo da adequabilidade da distribuição exponencial, se mantém como sendo uma distribuição qui-quadrado com $(k-1)$ graus de liberdade.

O teste de ajustamento bayesiano não paramétrico de Berger e Guglielmi (2001), apresentado na secção 3.3.3, pressupõe dois modelos bayesianos: um modelo paramétrico e um modelo não paramétrico, tal que o primeiro esteja incorporado no segundo. Para o estudo da adequabilidade da distribuição exponencial, apresentam-se, de seguida,

os respectivos modelos.

O modelo bayesiano paramétrico é dado por

$$X_i|\lambda \overset{\text{iid}}{\sim} \text{Exp}(\lambda), \text{ para } i = 1, 2, \ldots, n,$$
$$\lambda \sim \text{Ga}(a, b)$$

e o modelo bayesiano não paramétrico é

$$X_1, X_2, \ldots, X_n|G \overset{\text{iid}}{\sim} G$$
$$G|\Pi, \mathcal{A}_\lambda \sim \text{MPT}_M(\Pi, \mathcal{A}_\lambda),$$
$$\lambda \sim \text{Ga}(a, b)$$

onde $\text{MPT}_M(\Pi, \mathcal{A}_\lambda)$ define uma distribuição *a priori* mistura finita de árvores de Pólya, com parâmetros $(\Pi, \mathcal{A}_\lambda)$ e M níveis pré-especificados. Utiliza-se, novamente, como distribuição *a priori* para o parâmetro λ, a distribuição *a priori* não informativa da família conjugada natural, $\text{Ga}(a, b)$, com $a, b \to 0$.

As partições binárias (fixas, que não dependem de λ) são dadas por

$$B_{\varepsilon_{1:m}} = \left\{ \left(F_{\hat{\lambda}}^{-1}\left(\frac{k-1}{2^m}\right), F_{\hat{\lambda}}^{-1}\left(\frac{k}{2^m}\right)\right]\right\}, \tag{4.2}$$

para $m = 1, 2, \ldots, M$ e $k = 1, 2, \ldots, 2^m$, onde $F_{\hat{\lambda}}^{-1}(\cdot)$ define o quantil da distribuição exponencial cujo valor do parâmetro é a respectiva estimativa de máxima verosimilhança, $\hat{\lambda} = 1/\bar{x}$, com $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$.

Os valores para os parâmetros $\alpha_{\varepsilon_{1:m}}(\lambda)$ são obtidos utilizando as expressões definidas em (3.13) e (3.14) que, neste caso, vão ser dadas por

$$\alpha_{\varepsilon_{1:m-1}0}(\lambda) = \eta^{-1}\rho(m)\left(\frac{F_\lambda(B_{\varepsilon_{1:m-1}0})}{F_\lambda(B_{\varepsilon_{1:m-1}1})}\right)^{1/2} \tag{4.3}$$

e

$$\alpha_{\varepsilon_{1:m-1}1}(\lambda) = \eta^{-1}\rho(m)\left(\frac{F_\lambda(B_{\varepsilon_{1:m-1}1})}{F_\lambda(B_{\varepsilon_{1:m-1}0})}\right)^{1/2}, \tag{4.4}$$

para $m = 1, 2, \ldots, M$.

O parâmetro de escala η tem um papel fulcral no cálculo das estimativas do factor de Bayes, como já foi referido na secção 3.3.3. Opta-se por definir novamente $\eta = 2^s$, com s a tomar todos os valores inteiros dentro do intervalo $[-6, 6]$ e fixa-se, neste estudo, $\rho(m) = 4^m$.

4.2.1 Exemplos de aplicação

Com o objectivo de averiguar o comportamento dos testes bayesianos apresentados, para o estudo da adequabilidade da distribuição exponencial, utilizam-se dois conjuntos de dados simulados. Um dos conjuntos deve verificar a suposição de exponencialidade e o outro não deve verificar essa suposição. Desta forma, o primeiro conjunto de dados foi obtido simulando uma amostra de dimensão $n = 100$ de uma distribuição exponencial, Exp(1/5). O segundo conjunto de dados foi obtido simulando uma amostra de dimensão $n = 100$ de uma distribuição gama, Ga$(2, 1)$.

Na Figura 4.2, à esquerda, representa-se o histograma correspondente à amostra simulada de uma distribuição exponencial, Exp(1/5), com sobreposição das funções densidade de probabilidade exacta e estimada. Do lado direito representa-se o gráfico dos quantis empíricos contra os quantis teóricos.

Na Figura 4.3, à esquerda, representa-se o histograma correspondente à amostra simulada de uma distribuição gama, Ga(2,1), com sobreposição da função densidade de probabilidade exacta e da função densidade estimada, esta pressupondo que a amostra é proveniente de uma distribuição exponencial. Do lado direito representa-se o gráfico dos quantis empíricos contra os quantis teóricos (pressupondo exponencialidade).

Seguidamente, a cada um dos dois conjunto de dados simulados, são aplicados os dois testes bayesianos.

Para o cálculo das estimativas do factor de Bayes (ou do seu logaritmo) adapta-se o algoritmo 3, construindo o algoritmo 4, para a situação em estudo.

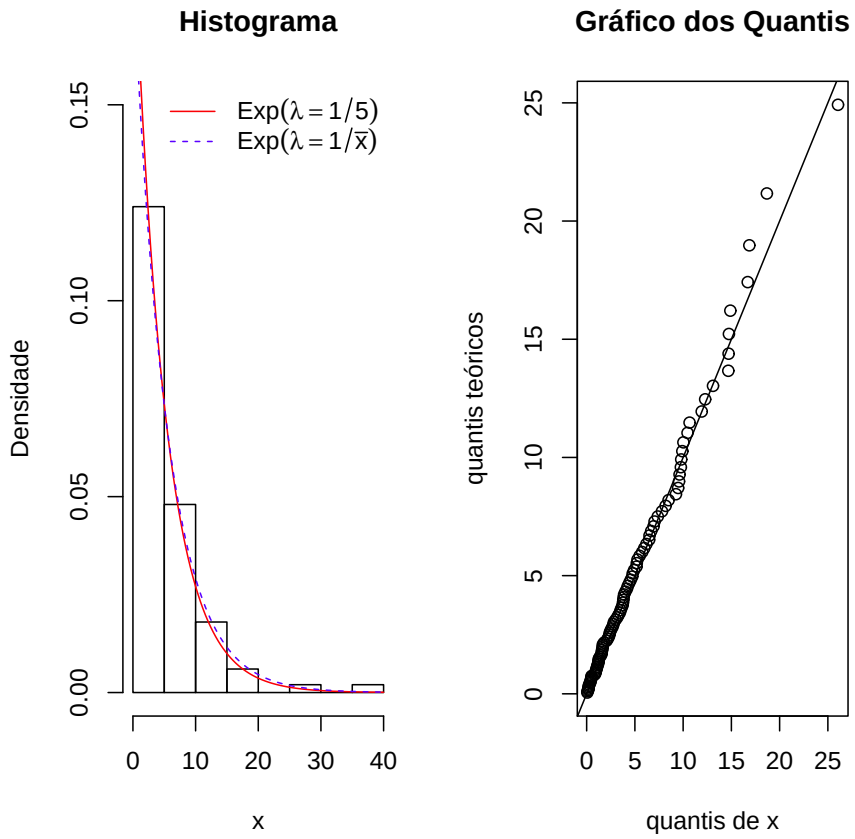

Figura 4.2: Histograma correspondente a uma amostra de dimensão $n = 100$ simulada de uma distribuição Exp(1/5), com sobreposição das funções densidade teórica e densidade estimada (à esquerda) e gráfico dos quantis empíricos contra os quantis teóricos (à direita).

Algoritmo 4

1. Define-se o número M de níveis da árvore (por exemplo, utilizando a sugestão de Hanson e Johnson (2002), $M \simeq \log_2(n)$);

2. Calculam-se as partições binárias do espaço amostral, $\mathbb{R}^+$, para os M níveis, definidas em (4.2);

3. Considera-se $\rho(m) = 4^m$ e $\eta = 2^s$, com s a tomar todos os valores inteiros no $[-6, 6]$;

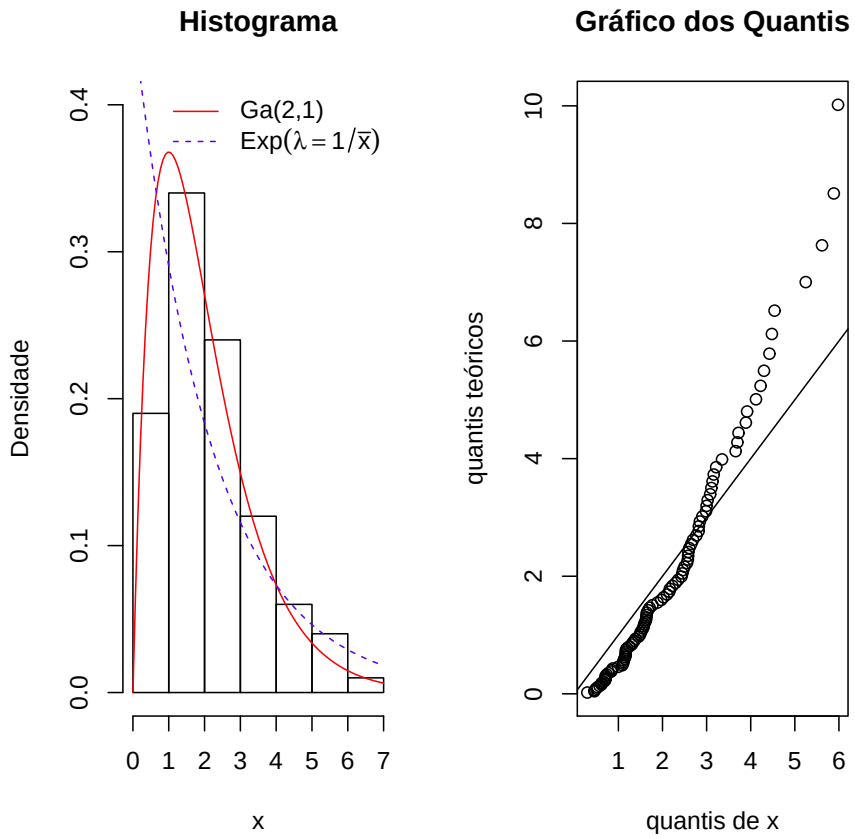

Figura 4.3: Histograma com sobreposição das funções densidade teórica e densidade estimada (esquerda) e gráfico dos quantis (direita) de uma amostra de dimensão $n = 100$, simulada de uma distribuição gama, $\text{Ga}(2, 1)$.

4. Para cada valor de η,

   (a) Para $l = 1, 2, \ldots, L$

      i. Gera-se um valor para o parâmetro, λ_l, da distribuição *a posteriori*;

      ii. Calculam-se os valores para os parâmetros da distribuição beta, dados por (4.3) e (4.4);

   (b) Calcula-se uma estimativa do factor de Bayes utilizando a expressão definida em (3.18);

5. Determina-se o valor mínimo das 13 estimativas calculadas do factor de Bayes.

Consideram-se, $M = 8$ níveis na árvore de Pólya, uma vez que a partir deste número de níveis não se observam diferenças significativas nas estimativas do factor de Bayes. É necessário utilizar $L = 20000$ valores simulados da distribuição *a posteriori* gama, $\text{Ga}(A, B)$, com $A = a + n\bar{x}$ e $B = b + n$, para garantir um erro padrão estimado de Monte Carlo pequeno.

Nas Figuras 4.4 e 4.5 encontram-se as representações gráficas das 13 estimativas do logaritmo do factor de Bayes assim como se indicam os valores mínimos do factor de Bayes e do seu logaritmo, para cada um dos dois conjuntos de dados em estudo, respectivamente.

Verifica-se que para os dados simulados da distribuição exponencial, o valor mínimo da estimativa do factor de Bayes é aproximadamente 1.07, ou seja, confirma-se que a distribuição exponencial se adequa ao conjunto de dados. Relativamente aos dados simulados da distribuição gama, o valor mínimo da estimativa do factor de Bayes é, agora, de aproximadamente 0.08. Neste caso, o teste rejeita a hipótese da distribuição exponencial ser o modelo adequado ao conjunto de dados, tal como era esperado.

Relativamente ao teste do qui-quadrado bayesiano, em vez de se simular um único valor da distribuição *a posteriori* de λ, e obter o correspondente valor da estatística de teste definida em (4.1), opta-se por simular 10000 valores da distribuição *a posteriori*. Utiliza-se $k = 6$ classes ($k \simeq 100^{0.4}$). A Figura 4.6 apresenta o histograma dos 10000 valores da estatística de teste, para a amostra simulada de uma distribuição exponencial, Exp(1/5), dos quais apenas 0.09% dos valores excedem o valor crítico,$\chi^2_{0.95}(5) = 11.0705$, a um nível de significância de 5%. Ou seja, apenas 9 das 10000 estatísticas de teste calculadas levam à rejeição da hipótese da distribuição exponencial se adequar ao conjunto de dados. Para a amostra simulada de uma distribuição gama, Ga(2,1), faz-se exactamente o mesmo estudo. A Figura 4.7 apresenta o histograma dos 10000 valores da estatística de teste, dos quais a totalidade dos valores excedem o valor crítico a um nível de significância de 5%. Neste caso, rejeita-se sempre a hipótese da distribuição exponencial ser a distribuição adequada para os dados que constituem esta amostra.

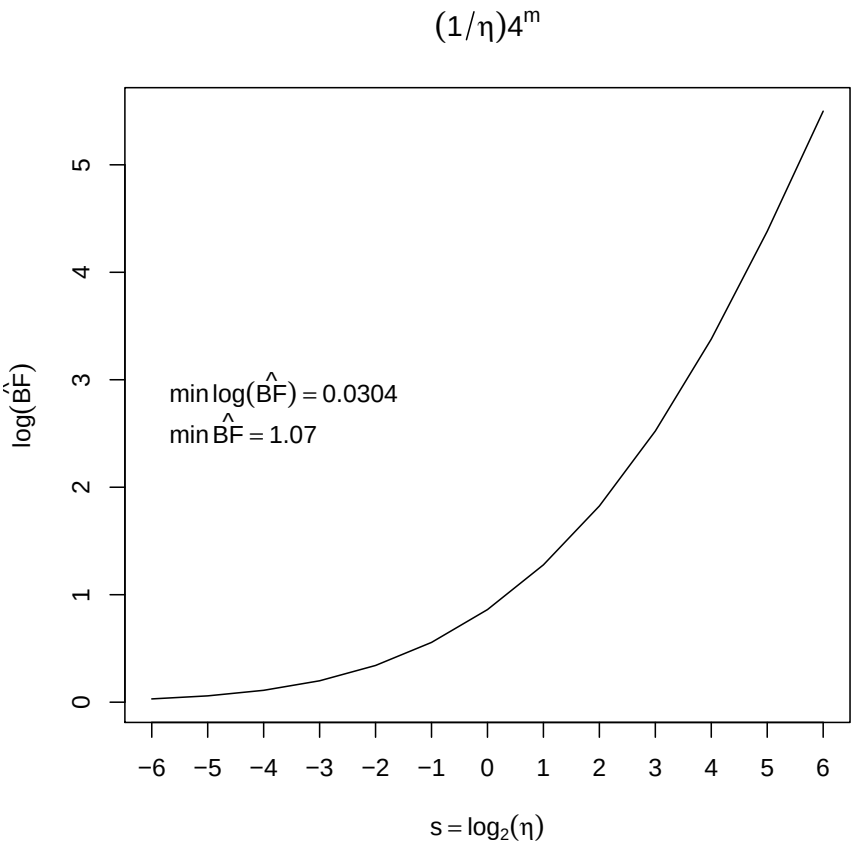

Figura 4.4: Estimativas do logaritmo do factor de Bayes para diferentes valores de $s = \log_2(\eta)$, para a amostra simulada de uma distribuição exponencial, Exp(1/5).

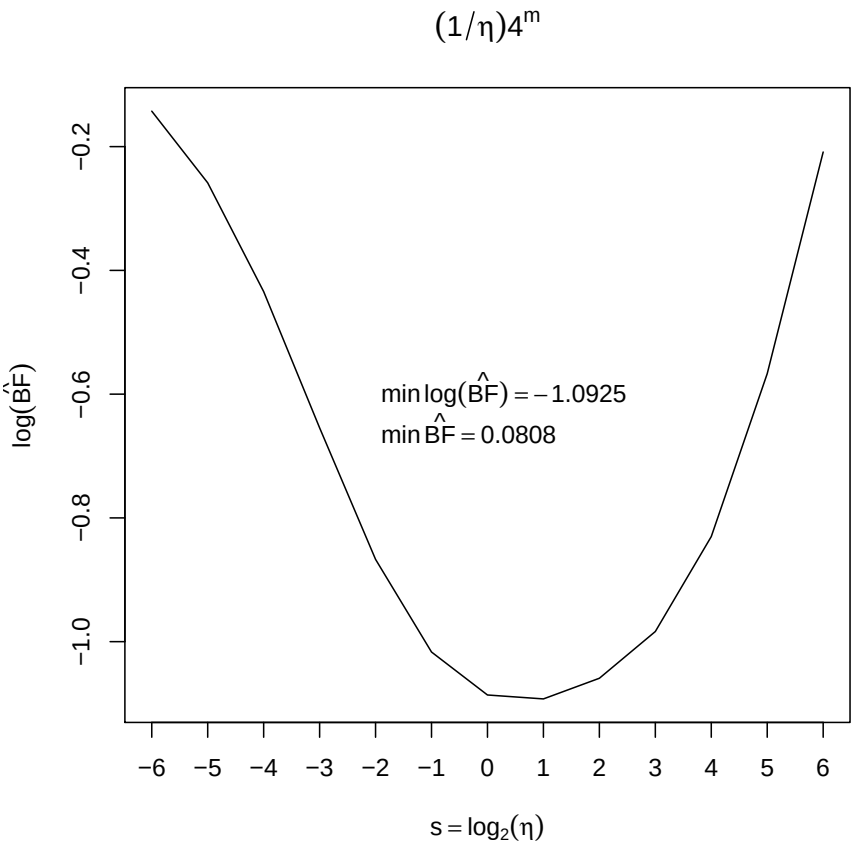

Figura 4.5: Estimativas do logaritmo do factor de Bayes para diferentes valores de $s = \log_2(\eta)$, para a amostra simulada de uma distribuição gama, Ga(2,1).

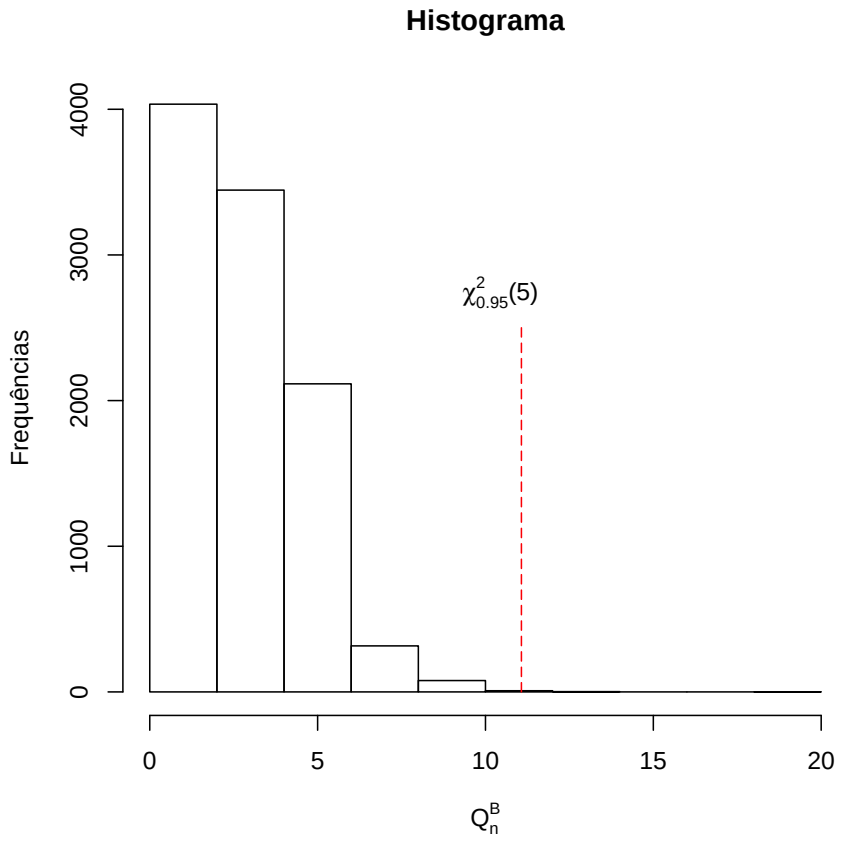

Figura 4.6: Histograma dos 10000 valores de Q_n^B para a amostra simulada de uma distribuição exponencial, Exp(1/5).

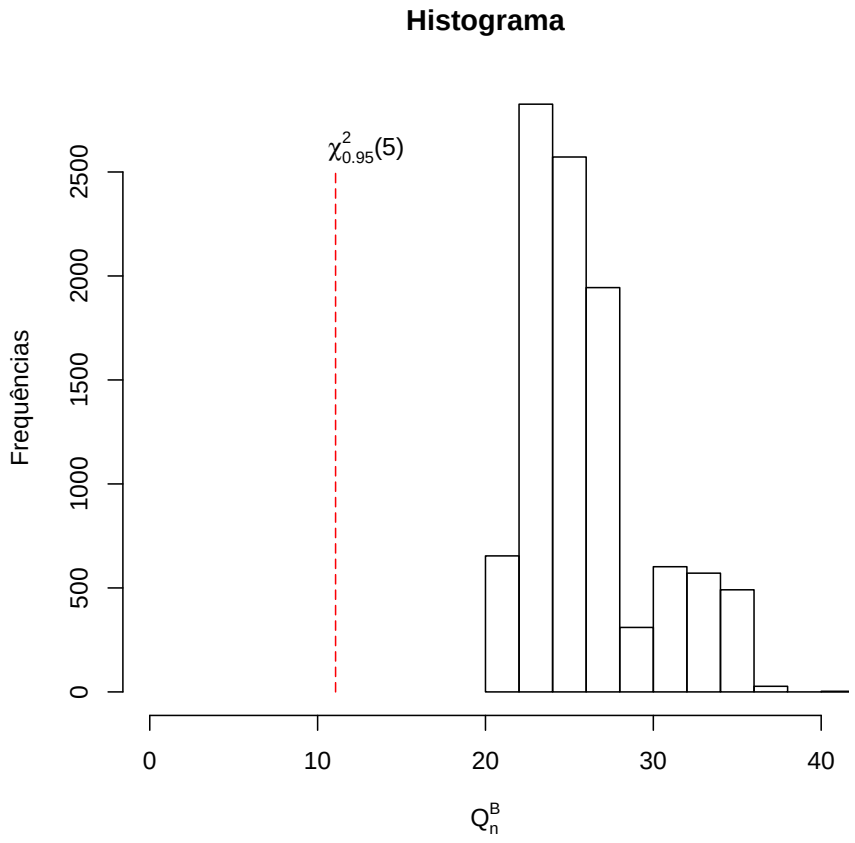

Figura 4.7: Histograma dos 10000 valores de Q_n^B para a amostra simulada de uma distribuição gama, Ga(2,1).

4.3 Estudo de simulação

Com o objectivo de investigar o desempenho dos diferentes testes propostos para o estudo da adequabilidade da distribuição exponencial (clássicos e bayesianos), realiza-se um estudo de simulação Monte Carlo.

A simulação Monte Carlo é um método muito utilizado para avaliar, empiricamente, o comportamento de um teste relativamente à taxa de erro tipo I e à sua potência. Desta forma, as simulações são divididas em duas partes. Na primeira parte, são simuladas amostras supondo H_0 verdadeira (utiliza-se a distribuição exponencial padrão). Caso seja rejeitada a hipótese nula então é cometido um erro de tipo I. Na segunda parte, são simuladas amostras supondo H_0 falsa (utilizam-se distribuições alternativas à exponencial, mas não exponenciais). Neste caso, se a hipótese é rejeitada, toma-se uma decisão correcta. As estimativas empíricas, para a taxa de erro tipo I e para a potência, são calculadas através da proporção de vezes que a hipótese nula é rejeitada, em cada uma das partes, respectivamente, com base num determinado número de amostras simuladas.

As distribuições alternativas à distribuição exponencial utilizadas neste estudo estão sumariadas na Tabela 4.4. Tenta-se abranger algumas distribuições frequentemente consideradas em outros estudos, bem como ter uma variedade de distribuições com diferentes funções taxa de falha (crescente, decrescente e não monótona). Por exemplo, as distribuições gama e Weibull estão associadas a funções taxa de falha crescentes quando $a > 1$ e estão associadas a funções taxa de falha decrescentes quando $0 < a < 1$. Para $a = 1$, ambas as distribuições anteriores coincidem com a distribuição exponencial padrão com função taxa de falha constante. A função taxa de falha para a distribuição lognormal, inicialmente cresce com o tempo para depois decrescer (não-monótona). A distribuição Half-Normal tem função taxa de falha crescente e a distribuição $\chi^2(1)$ tem função taxa de falha decrescente. A distribuição Half-Cauchy caracteriza-se por ter uma cauda pesada.

Tabela 4.4: Distribuições alternativas à distribuição exponencial.

Distribuição	Notação	Densidade
Gama	$\text{Ga}(a,1)$	$\Gamma(a)^{-1}x^{a-1}\exp(-x)$
Weibull	$\text{Wei}(a,1)$	$ax^{a-1}\exp(-x^a)$
LogNormal	$\text{LN}(0,1)$	$x^{-1}(2\pi)^{-1/2}\exp(-\log^2(x)/2)$
Half-Normal	$\text{HN}(0,1)$	$(2/\pi)^{1/2}\exp(-x^2/2)$
Qui-quadrado	$\chi^2(a)$	$1/(2^{a/2}\Gamma(a/2))x^{a/2-1}\exp(-x/2)$
Half-Cauchy	$\text{HCa}(0,1)$	$(2/\pi)(x^2+1)$

A escolha dos parâmetros das diferentes distribuições alternativas é feita de modo que a densidade resultante se afaste gradualmente da forma de uma distribuição exponencial padrão. Todas as distribuições estão representadas graficamente na Figura 4.8, assim como também se representa graficamente a distribuição exponencial padrão, com o objectivo de permitir comparar a curva da sua densidade com a de todas as outras densidades.

De acordo com a filosofia dos testes clássicos, procura-se um teste que mantenha a taxa de erro tipo I próxima de um dado nível de significância fixo, α, e que tenha a maior potência possível. No entanto, o teste de ajustamento bayesiano não paramétrico não está construído para controlar qualquer tipo de erro. Consequentemente, é necessário construir uma regra ou região de rejeição para o teste de ajustamento bayesiano, por forma a que seja possível a comparação com os diferentes testes apresentados.

Na primeira parte deste estudo de simulação, simulam-se 500 amostras, de três dimensões diferentes ($n = 25$, $n = 50$ e $n = 100$), da distribuição exponencial padrão.

O factor de Bayes é estimado de acordo com o algoritmo 4. Relativamente ao número de níveis, M, a utilizar na construção da árvore de Pólya para cada uma das três dimensões, realiza-se previamente um pequeno estudo, onde se comparam as

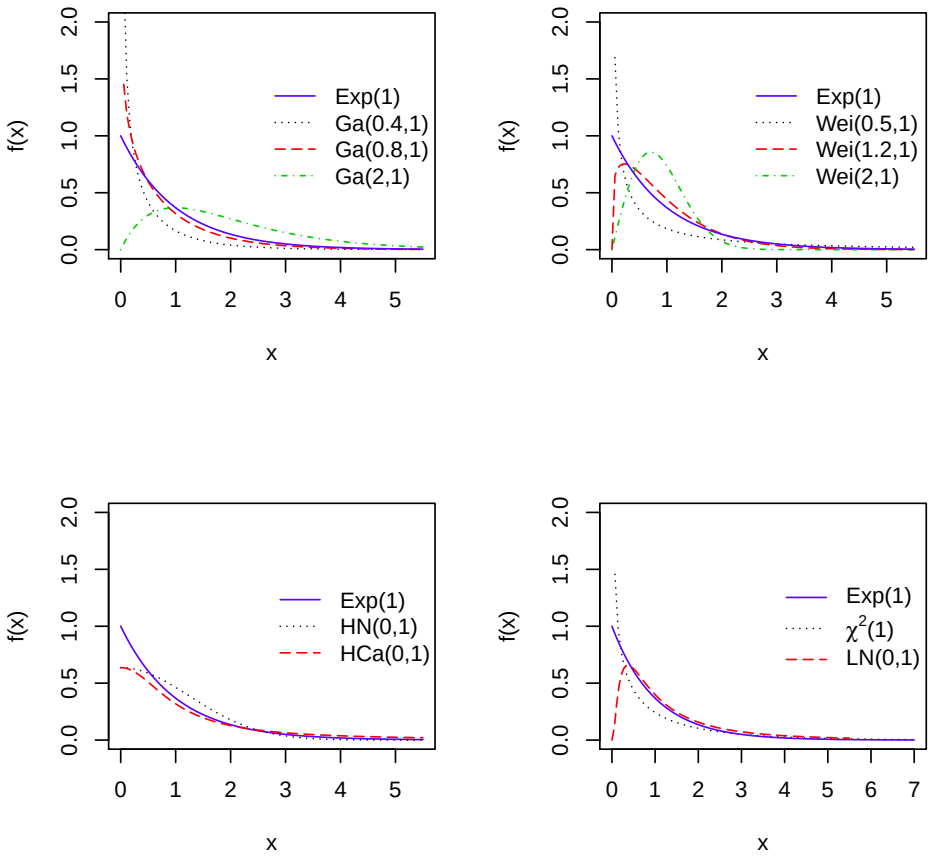

Figura 4.8: Representação gráfica de algumas distribuições alternativas à distribuição exponencial.

estimativas do factor de Bayes à medida que se aumenta o número de níveis. De acordo com a regra $M \simeq \log_2(n)$, variam-se os valores de M entre 5 e 10. Para as três dimensões amostrais utilizadas tem-se que para $M \geq 8$ não se verificam diferenças significativas nos resultados e opta-se, assim, por fixar para todas as dimensões amostrais uma árvore de Pólya com $M = 8$ níveis. Utilizam-se novamente $L = 20000$ valores simulados da distribuição *a posteriori* para o parâmetro λ.

A estimativa do factor de Bayes é definida a favor do modelo paramétrico (H_0) e contra o modelo não paramétrico (H_1), $\widehat{\mathrm{BF}}_{01}(x)$. Portanto, a região de rejeição do teste pode ser construída como $\{x : \widehat{\mathrm{BF}}_{01}(x) < c_{BF}\}$, onde $c_{BF} > 0$ vai ser designado por *threshold* crítico empírico para a estimativa do factor de Bayes.

Para obter o *threshold* crítico utiliza-se o mesmo procedimento aplicado ao cálculo do valor crítico frequencista, isto é, com base nos resultados obtidos por simulação Monte Carlo. Desta forma, utilizam-se as 500 amostras simuladas da distribuição exponencial padrão e calcula-se a estimativa do factor de Bayes para cada uma das amostras. O *threshold* crítico empírico para o factor de Bayes, e para cada dimensão amostral, é determinado a partir do quantil $100\alpha\%$ da correspondente função de distribuição empírica.

Na Tabela 4.5 apresenta-se o *threshold* crítico empírico para a estimativa do factor de Bayes, c_{BF}, necessário para atingir uma taxa de erro tipo I de 5% para cada uma das dimensões amostrais consideradas.

Tabela 4.5: *Threshold* crítico empírico para $\widehat{\mathrm{BF}}_{01}(x)$.

n	25	50	100
c_{BF}	0.4876	0.5697	0.7117

Para o teste do qui-quadrado bayesiano, simula-se um único valor da distribuição *a posteriori* de λ e calcula-se o valor da estatística de teste definida em (4.1). Neste

teste, é necessário definir o número de classes de acordo com a dimensão da amostra. Johnson (2004) recomenda a utilização da regra $k \simeq n^{0.4}$. Também, neste caso, se varia o número de classes, mas os resultados não se alteram significativamente. No estudo de simulação considera-se $k = 4$, 5 e 6, para $n = 25$, 50 e 100, respectivamente.

Na segunda parte deste estudo de simulação, simulam-se 500 amostras, de três dimensões diferentes ($n = 25$, $n = 50$ e $n = 100$) das distribuições alternativas à distribuição exponencial, Ga(0.4,1), Ga(0.8,1), Ga(2,1), Wei(0.5,1), Wei(1.2,1), Wei(2,1), HN(0,1), HCa(0,1), LN(0,1) e $\chi^2(1)$, respectivamente.

Para cada amostra simulada em cada uma das duas partes, realizam-se cada um dos oito testes (2 bayesianos e 6 clássicos). Considera-se para nível de significância $\alpha = 5\%$ e verifica-se se a hipótese nula é ou não rejeitada.

4.3.1 Resultados e discussão

Nas Tabelas 4.6 e 4.7 apresentam-se a média (e o desvio-padrão) das estimativas empíricas da taxa de erro tipo I e da potência de cada um dos testes, em função da dimensão da amostra.

Nos casos em que a hipótese nula é falsa, avalia-se a potência dos diferentes testes. O teste do qui-quadrado bayesiano é o que teve o pior desempenho, independentemente da dimensão da amostra, sendo portanto não recomendável.

Para as distribuições alternativas com função taxa de falha crescente (Ga(2,1), Wei(1.2,1), Wei(2,1) e HN(0,1)), a potência empírica do teste de ajustamento bayesiano é quase sempre superior à dos testes clássicos. Apenas nas distribuições Wei(1.2,1) e HN(0,1) e para dimensões amostrais mais pequenas, a estatística $T_{n,a}$ é ligeiramente mais potente. No entanto, esta última estatística é para todas as outras distribuições alternativas consideradas a estatística clássica menos potente.

Quando as amostras simuladas são de distribuições alternativas com função taxa de

Tabela 4.6: Média (e desvio padrão) da estimativa empírica para a proporção de rejeições correctas para cada um dos testes. Para a distribuição Exp(1), tem-se a taxa de erro tipo I e para todas as outras distribuições a respectiva potência do teste.

		Teste										
Distribuição	n	$\widehat{\mathrm{BF}}_{01}(x)$	$Q_n^B(\check{\theta})$	EP_n	CO_n	$\overline{\mathrm{CM}}_n$	AD_n	$\mathrm{BH}_{n,a=1}$	$\mathrm{BH}_{n,a=1.5}$	$\mathrm{BH}_{n,a=2.5}$	$\mathrm{T}_{n,a=1.5}$	$\mathrm{T}_{n,a=2.5}$
Exp(1)	25	0.050	0.052	0.038	0.034	0.048	0.046	0.042	0.042	0.046	0.048	0.046
		(0.028)	(0.028)	(0.022)	(0.021)	(0.030)	(0.023)	(0.024)	(0.027)	(0.025)	(0.028)	(0.036)
	50	0.050	0.052	0.034	0.042	0.044	0.052	0.044	0.040	0.042	0.05	0.052
		(0.030)	(0.041)	(0.017)	(0.031)	(0.023)	(0.027)	(0.026)	(0.021)	(0.022)	(0.017)	(0.023)
	100	0.050	0.052	0.056	0.050	0.050	0.048	0.054	0.056	0.054	–	–
		(0.031)	(0.043)	(0.032)	(0.033)	(0.027)	(0.037)	(0.031)	(0.039)	(0.033)	–	–
Ga(0.4, 1)	25	0.898	0.404	0.820	0.956	0.824	0.942	0.886	0.868	0.844	0.302	0.498
		(0.040)	(0.084)	(0.042)	(0.025)	(0.042)	(0.026)	(0.041)	(0.045)	(0.046)	(0.061)	(0.061)
	50	0.990	0.856	0.984	0.998	0.984	0.998	0.996	0.992	0.986	0.882	0.918
		(0.011)	(0.037)	(0.021)	(0.006)	(0.021)	(0.006)	(0.008)	(0.014)	(0.019)	(0.022)	(0.017)
	100	1	1	1	1	1	1	1	1	1	–	–
Ga(0.8, 1)	25	0.110	0.042	0.114	0.148	0.110	0.146	0.136	0.130	0.126	0.004	0.026
		(0.046)	(0.024)	(0.049)	(0.052)	(0.044)	(0.052)	(0.047)	(0.052)	(0.046)	(0.008)	(0.019)
	50	0.214	0.080	0.212	0.266	0.202	0.248	0.232	0.216	0.222	0.05	0.092
		(0.071)	(0.048)	(0.067)	(0.068)	(0.068)	(0.078)	(0.077)	(0.076)	(0.068)	(0.029)	(0.037)
	100	0.318	0.310	0.348	0.452	0.328	0.392	0.380	0.372	0.348	–	–
		(0.083)	(0.054)	(0.063)	(0.095)	(0.079)	(0.093)	(0.074)	(0.074)	(0.058)	–	–
Ga(2, 1)	25	0.696	0.310	0.588	0.578	0.592	0.554	0.650	0.638	0.612	0.674	0.676
		(0.060)	(0.054)	(0.088)	(0.066)	(0.088)	(0.076)	(0.077)	(0.076)	(0.082)	(0.088)	(0.092)
	50	0.989	0.630	0.924	0.984	0.918	0.928	0.956	0.942	0.928	0.936	0.912
		(0.019)	(0.064)	(0.025)	(0.025)	(0.030)	(0.021)	(0.033)	(0.022)	(0.023)	(0.036)	(0.045)
	100	1	0.932	0.996	0.998	0.996	0.998	0.998	0.998	0.996	–	–
			(0.035)	(0.008)	(0.006)	(0.008)	(0.006)	(0.006)	(0.006)	(0.008)	–	–
$\chi^2(1)$	25	0.576	0.25	0.626	0.810	0.614	0.782	0.704	0.672	0.656	0.130	0.250
		(0.068)	(0.057)	(0.067)	(0.043)	(0.075)	(0.049)	(0.079)	(0.077)	(0.078)	(0.045)	(0.054)
	50	0.880	0.592	0.882	0.982	0.872	0.962	0.926	0.910	0.890	0.674	0.738
		(0.034)	(0.074)	(0.033)	(0.020)	(0.025)	(0.028)	(0.028)	(0.025)	(0.027)	(0.061)	(0.031)
	100	1	0.928	0.992	1	0.988	1	1	0.998	0.992	–	–
			(0.030)	(0.014)		(0.017)			(0.006)	(0.014)	–	–

Tabela 4.7: Média (e desvio padrão) da estimativa empírica para a proporção de rejeições correctas para cada um dos testes. Para a distribuição Exp(1), tem-se a taxa de erro tipo I e para todas as outras distribuições a respectiva potência do teste (continuação).

		Teste										
Distribuição	n	$\widehat{\mathrm{BF}}_{01}(x)$	$Q_n^B(\check{\theta})$	EP_n	CO_n	$\widehat{\mathrm{CM}}_n$	AD_n	$\mathrm{BH}_{n,a=1}$	$\mathrm{BH}_{n,a=1.5}$	$\mathrm{BH}_{n,a=2.5}$	$\mathrm{T}_{n,a=1.5}$	$\mathrm{T}_{n,a=2.5}$
Wei$(0.5,1)$	25	0.990	0.664	0.938	0.976	0.940	0.974	0.966	0.954	0.944	0.452	0.706
		(0.024)	(0.071)	(0.037)	(0.025)	(0.035)	(0.019)	(0.023)	(0.027)	(0.036)	(0.103)	(0.068)
	50	1	0.986	0.998	1	1	1	1	1	1	0.974	0.998
			(0.019)	(0.006)	1	1	(0.019)	(0.010)				
	100	1	1	1	1	1	1	1	1	1	–	–
Wei$(1.2,1)$	25	0.208	0.084	0.150	0.124	0.164	0.144	0.164	0.160	0.158	0.226	0.214
		(0.040)	(0.049)	(0.041)	(0.048)	(0.041)	(0.047)	(0.042)	(0.036)	(0.045)	(0.070)	(0.072)
	50	0.294	0.132	0.270	0.260	0.276	0.224	0.242	0.284	0.288	0.312	0.318
		(0.038)	(0.065)	(0.039)	(0.057)	(0.056)	(0.053)	(0.071)	(0.070)	(0.047)	(0.061)	(0.058)
	100	0.635	0.242	0.598	0.594	0.574	0.530	0.612	0.606	0.594	–	–
		(0.029)	(0.038)	(0.033)	(0.057)	(0.042)	(0.024)	(0.044)	(0.040)	(0.033)	–	–
Wei$(2,1)$	25	0.990	0.742	0.982	0.972	0.980	0.978	0.970	0.984	0.984	0.988	0.988
		(0.019)	(0.045)	(0.020)	(0.021)	(0.019)	(0.017)	(0.017)	(0.016)	(0.016)	(0.017)	(0.017)
	50	1	0.994	1	1	1	1	1	1	1	1	1
			(0.009)									
	100	1	1	1	1	1	1	1	1	1	–	–
HCa$(0,1)$	25	0.766	0.514	0.764	0.742	0.770	0.754	0.758	0.768	0.782	0.038	0.250
		(0.030)	(0.059)	(0.039)	(0.033)	(0.036)	(0.038)	(0.035)	(0.036)	(0.036)	(0.037)	(0.058)
	50	0.968	0.752	0.950	0.928	0.956	0.936	0.946	0.948	0.958	0.346	0.564
		(0.020)	(0.039)	(0.025)	(0.030)	(0.023)	(0.025)	(0.021)	(0.025)	(0.024)	(0.041)	(0.059)
	100	0.998	0.958	0.998	0.998	0.998	0.998	0.998	1	1	–	–
		(0.004)	(0.017)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)			–	–
LN$(0,1)$	25	0.270	0.114	0.134	0.090	0.170	0.170	0.122	0.144	0.168	0.076	0.056
		(0.038)	(0.031)	(0.049)	(0.037)	(0.040)	(0.052)	(0.050)	(0.061)	(0.053)	(0.028)	(0.021)
	50	0.352	0.210	0.168	0.140	0.266	0.342	0.206	0.204	0.216	0.094	0.072
		(0.042)	(0.057)	(0.067)	(0.057)	(0.075)	(0.078)	(0.057)	(0.065)	(0.063)	(0.034)	(0.039)
	100	0.740	0.446	0.234	0.184	0.446	0.704	0.296	0.294	0.302	–	–
		(0.038)	(0.064)	(0.048)	(0.063)	(0.074)	(0.047)	(0.062)	(0.046)	(0.061)	–	–
HN$(0,1)$	25	0.267	0.118	0.232	0.160	0.246	0.198	0.246	0.254	0.256	0.314	0.322
		(0.036)	(0.054)	(0.042)	(0.056)	(0.041)	(0.044)	(0.037)	(0.031)	(0.035)	(0.036)	(0.046)
	50	0.586	0.192	0.500	0.372	0.514	0.408	0.462	0.480	0.506	0.576	0.586
		(0.040)	(0.065)	(0.085)	(0.079)	(0.071)	(0.073)	(0.087)	(0.077)	(0.082)	(0.069)	(0.060)
	100	0.866	0.398	0.848	0.722	0.864	0.792	0.810	0.834	0.850	–	–
		(0.038)	(0.043)	(0.048)	(0.068)	(0.044)	(0.057)	(0.061)	(0.057)	(0.044)	–	–

falha decrescente, como é o caso das distribuições Ga(0.4,1) e Wei (0.5,1), o teste de ajustamento bayesiano é, pelo menos, tão potente quanto os testes clássicos. Contudo, é ligeiramente menos potente do que alguns dos testes clássicos para as restantes distribuições com função taxa de falha decrescente (Ga(0.8,1) e $\chi^2(1)$), talvez por estas duas distribuições estarem mais próximas de uma distribuição exponencial padrão.

Para a distribuição Half-Cauchy, a potência empírica do teste de ajustamento bayesiano é comparável com a dos testes clássicos e para a distribuição lognormal, particularmente quando as amostras são de pequena dimensão, o teste de ajustamento bayesiano é o que apresenta melhor desempenho. Quando as amostras simuladas provêm de distribuições cujos parâmetros impliquem um afastamento gradual da distribuição exponencial, os valores da potência empírica atingem valores elevados (> 0.80) e, por vezes, iguais a 1, para dimensões amostrais elevadas ($n = 100$). Os testes clássicos mantêm a instabilidade, em termos de potência, já mencionada nos estudos de Henze e Meintanis (2005), isto é, não existe um teste mais potente que todos os outros. O mesmo se verifica para o teste de Anderson-Darling, sendo, em algumas situações, este teste superior, em termos de potência, aos seus concorrentes clássicos, mas muito próximo dos resultados obtidos com o uso da estatística de Cox e Oakes, CO_n. Para as distribuições alternativas consideradas, o número de amostras rejeitadas correctamente aumenta, como seria de esperar, à medida que a dimensão de cada amostra aumenta, independentemente do teste (bayesiano ou clássico).

Capítulo 5

Conclusões e discussão

O estudo da adequabilidade do modelo probabilístico proposto para representar o fenómeno aleatório que produz um conjunto de dados é fundamental para muitos dos procedimentos estatísticos.

A abordagem clássica tem sido amplamente utilizada na procura de novos métodos para o estudo da adequabilidade de um modelo, mas, apesar do grande número de trabalhos publicados, não existe e, provavelmente nunca existirá um método que consiga captar simultaneamente todos os possíveis desvios ao modelo pressuposto. Relativamente à abordagem bayesiana, os métodos existentes ainda são muito escassos e embora se diferenciem bastante na sua construção, relativamente à abordagem clássica, o seu estudo ainda representa um grande desafio de investigação.

O principal objectivo deste trabalho é o de apresentar um teste bayesiano não paramétrico para o estudo da adequabilidade da distribuição exponencial que, comparativamente com os testes clássicos existentes, tenha um melhor desempenho em termos de potência.

Sendo o teste bayesiano não paramétrico um método que requer a definição de um modelo bayesiano não paramétrico, dedicou-se parte deste trabalho à apresentação de alguns modelos bayesianos não paramétricos usuais, dando particular destaque e

desenvolvimento à distribuição árvore de Pólya.

Neste trabalho, apresentaram-se ainda os métodos bayesianos existentes para o estudo da adequabilidade de um modelo, quer para variáveis discretas quer para variáveis contínuas, em particular, o valor-p preditivo, o teste do qui-quadrado bayesiano e os testes de ajustamento bayesianos.

No caso do estudo da adequabilidade de um modelo para variáveis discretas propôs-se o factor pseudo-Bayes como método alternativo aos métodos bayesianos existentes. Fez-se, neste caso, um pequeno estudo comparativo sobre a adequabilidade do modelo Poisson a um conjunto de dados. Neste estudo, simularam-se amostras do modelo Poisson e de outros modelos para dados discretos, como, por exemplo, o modelo binomial e o modelo binomial negativo. Seguidamente, calculou-se o factor pseudo-Bayes, o factor de Bayes fraccionário, o factor de Bayes fraccionário hierárquico e o valor-p de discrepância.

Quando a amostra é simulada de uma distribuição de Poisson, qualquer um dos métodos considerados validou o modelo. Porém, é de notar que quando se considera o factor de Bayes fraccionário os resultados dependem do parâmetro de concentração c: o modelo é, em regra, validado quando este parâmetro assume valores pequenos. Quando a amostra é simulada de uma das distribuições alternativas, o factor de Bayes fraccionário hierárquico revela-se instável, dificultando a decisão de validar ou não o modelo de Poisson. Quando a amostra é simulada a partir de uma distribuição alternativa significativamente afastada da distribuição de Poisson, todos os métodos rejeitam correctamente a hipótese de adequabilidade dos dados ao modelo Poissoniano, sendo, contudo, de notar o melhor desempenho por parte da estimativa do valor-p de discrepância, bem como do factor pseudo-Bayes. Note-se, ainda, que quando a amostra é simulada de uma distribuição próxima do modelo de Poisson, nenhum dos métodos rejeita o referido modelo. Quando a amostra é proveniente de uma população ligeiramente afastada de uma população de Poisson, somente o factor pseudo-Bayes e a estimativa do valor-p mostram a inadequabilidade dos dados ao modelo. Assim, em conclusão,

pode dizer-se que o estudo de simulação efectuado (os detalhes foram apresentados na secção 3.3.2) aponta para um bom desempenho do factor pseudo-Bayes, relativamente aos restantes métodos bayesianos utilizados. No entanto, será interessante fazer para trabalho futuro, um estudo de simulação para a análise da adequação de modelos para dados discretos incluindo estes métodos e os métodos clássicos.

Foi, também, realizado um breve estudo considerando dados contínuos. Para isso, considerou-se uma amostra proveniente de uma população normal e uma outra amostra com origem numa população não normal. Verificou-se que o teste de ajustamento bayesiano não paramétrico não rejeita a hipótese de normalidade no primeiro caso e rejeita essa mesma hipótese no segundo caso. Relativamente à utilização das duas propostas para definir o valor de η (os detalhes são apresentados na secção 3.3), notou-se que não há diferenças significativas nos valores obtidos quando o modelo suposto é rejeitado. Contudo, quando a distribuição hipotética se adequa aos dados, valores de η inferiores à unidade e próximos de zero conduzem a um factor de Bayes com um valor próximo de um. Este resultado pode ter grande interesse quando, num estudo de simulação, é necessário definir um valor de corte (*threshold*) para o factor de Bayes.

Relativamente ao estudo da adequabilidade do modelo exponencial, objectivo principal deste trabalho, são propostos dois testes bayesianos, o teste bayesiano não paramétrico de Berger e Guglielmi (2001) e o teste do qui-quadrado de Pearson, apresentado por Johnson (2004), fazendo as adaptações necessárias.

A literatura referente à abordagem clássica para o estudo da adequabilidade do modelo exponencial é bastante extensa. Um dos trabalhos mais recentes sobre o desempenho de muitos dos testes clássicos propostos na literatura é apresentado por Henze e Meintanis (2005). Com base no estudo de simulação dos referidos autores e nas conclusões obtidas, seleccionaram-se os cinco testes clássicos mais potentes e considerou-se, ainda, o teste clássico de Anderson e Darling (1954).

Com o objectivo de comparar o desempenho dos dois testes bayesianos propostos com o de alguns dos testes clássicos mais potentes, realizou-se um estudo de simu-

lação. Consideraram-se, neste estudo, como distribuições alternativas à distribuição exponencial, a distribuição Gama, a distribuição Weibull, a distribuição Log-Normal, a distribuição Half-Normal, a distribuição do Qui-Quadrado e a distibuição Half-Cauchy. Os parâmetros destas distribuições alternativas foram escolhidos de modo a que as formas das distribuições fossem diferindo da forma de uma distribuição exponencial padrão.

Para as distribuições alternativas com taxa de falha crescente, notou-se que a potência empírica do teste de ajustamento bayesiano é quase sempre superior à dos testes clássicos. Por outro lado, quando as amostras simuladas são obtidas a partir de distribuições alternativas com função taxa de falha decrescente, o teste de ajustamento bayesiano é, pelo menos, tão potente quanto os clássicos. Saliente-se, ainda, o facto de que quando as amostras são de pequena dimensão, o teste de ajustamento bayesiano é o que apresenta melhor desempenho. Relativamente aos testes clássicos, este revelaram uma grande instabilidade em termos de potência. Assim, pode afirmar-se que o estudo de simulação efectuado, não sendo exaustivo, na medida que se restringiu o trabalho a distribuições alternativas usualmente consideradas em outros estudos, permite concluir que o teste bayesiano não paramétrico proposto para o estudo da adequabilidade da distribuição exponencial tem, de uma forma geral, um bom desempenho. O mesmo já não se pode dizer do teste do qui-quadrado bayesiano que, em comparação com os testes clássicos, teve o pior desempenho.

Como trabalho futuro, pretende-se investigar a possibilidade de generalizar o teste bayesiano não paramétrico para as distribuições pertencentes à família exponencial.

Anexo A

Código em R

Seja $x = (x_1, x_2, \ldots, x_n)$ uma amostra observada. O código em linguagem R para a implementação do algoritmo 4 apresentado no Capítulo 4 é o seguinte.

```
################################################################
Programa Principal (último a correr)
################################################################
# Amostra observada
X<-c()
# Amostra observada ordenada
x<-sort(X)
#13 valores de eta=h
h.vec <- 2^(-6:6)
# Calcula para cada um dos 13 valores de h a função psi
pt.bf <- sapply(h.vec, function(a) pt.gof(x, h = a))
BF<-c()
for(i in 1:13){
    # Estimativa do factor de Bayes para cada valor de h
    BF[i]<- 1/mean((pt.bf[,i]))
```

```
    }
BF
min(BF)

#################################################################
Função Principal pt.gof
#################################################################
pt.gof<-function(x,h){
# Inicializações gerais
n <- length(x)
xb <-mean(x)
lambda<-1/xb
# Número de níveis da árvore
M <- 8
# Número de parâmetros simulados
L <- 20000
# Intervalos das partições
B <-interval.obs(M,xb)
# Números de observações e identificação da ordem
# das observações nos Intervalos das partições
B1 <- class.obs(B,x)[[1]]
B2 <- class.obs(B,x) [[2]]

# Ciclo para o cálculo do factor de Bayes
# L valores simulados da distribuição a posteriori para lambda

set.seed(12345)
lambda_sim<-c()
lambda_sim<-rgamma(L,shape=n,rate=n*xb)
```

```
psi <- c()
for (l in 1:L){
    lambda<-lambda_sim[l]
    P <- miuB_eps(B,lambda)
    C<-ceps(B,P)
    A<-alfaeps(B,C,h)
    lpsi<-logpsi_thetal(B,x,A,B1,B2)
    psi[l]<-exp(sum(lpsi))
    }
return(psi)
}

#################################################################
Funções Auxiliares (devem correr antes do Programa Principal)
#################################################################

#################################################################
FUNÇÃO 1 - Partições do espaço amostral: esta função calcula os
limites à direita dos sub-intervalos da árvore de Pólya com M níveis
centrada numa distribuição Exp(lamdba), lambda=Est. Máx. Veros.
#################################################################
interval.obs <- function(M,xb){
# Medida central   G0 = Exp(lambda)
lambda <- 1/xb
# Inicializa a lista de valores
B <- as.list(1:M)

for(m in 1:M){
    # Limites à direita dos sub-intervalos
```

```
    q <- 1/(2**m)*(1:(2**m-1))
    B[[m]] <- c(qexp(q,rate=lambda),25000)
    }
return(B)
}

#####################################################################
Função 2 - Ciclo para definir as listas B1 e B2 para auxilio no
cálculo da função psi
#####################################################################
class.obs<-function(B,x){
M <- length(B)
B1 <- as.list(1:M)
B2 <- as.list(1:M)

  for(m in 1:M){
    old_aux = c()
    pos_aux = c()
    for (k in 1:length(B[[m]])){
        v_aux=which(x<=B[[m]][k])
        list_ins = setdiff(v_aux,old_aux)
        if (length(list_ins)==0){
            list_ins=0
        }
        B1[[m]][k] = list(list_ins)
        pos_aux = c(pos_aux,rep(k,length(list_ins)))
        B2[[m]] = pos_aux
        old_aux=v_aux
    }
```

```
   }
return(list(B1,B2))
}

#################################################################
Função 3 - Calcula as medidas de probabilidade miuB_eps, armazenada
na lista P, para cada sub-intervalo em B
#################################################################
miuB_eps<-function(B,lambda){
M <- length(B)
P <- as.list(1:M)

for (m in 0:(M-1)){
    for(j in 1:2**m){
        j0 <- (j-1)*2 + 1
        j1 <- (j-1)*2 + 2
        P0<-pexp(B[[m+1]][j0],lambda)
        if (m>0) {
            if (j>1){
                if (j!=2**m) {
                    P00<-pexp(B[[m+1]][j0-1],lambda)
                    P1<-pexp(B[[m+1]][j1],lambda)
                    P[[m+1]][j0]<-P0-P00
                    P[[m+1]][j1]<-P1-P0
                    } else {
                        P00<-pexp(B[[m+1]][j0-1],lambda)
                        P[[m+1]][j0]<-P0-P00
                        P[[m+1]][j1]<-1-P0
                    }
```

```
                } else {
                P1<-pexp(B[[m+1]][j1],lambda)
                P[[m+1]][j0]<-P0
                P[[m+1]][j1]<-P1-P0
                }
            } else {
                P[[m+1]][j0]<-P0
                P[[m+1]][j1]<-1-P0
                }
    }
}
return(P)
}

#################################################################
Função 4 - Calcula os c_eps=miuB_eps1/miu_B_eps0 para cada
sub-intervalo em B
#################################################################
ceps<-function(B,P){
M <- length(B)
c_eps <- as.list(1:M)

for (m in 0:(M-1)){
    for(j in 1:2**m){
        j0 <- (j-1)*2 + 1
        j1 <- (j-1)*2 + 2
        c_eps[[m+1]][j0] <- exp(log(P[[m+1]][j1])-log(P[[m+1]][j0]))
        c_eps[[m+1]][j1] <- exp(log(P[[m+1]][j0])-log(P[[m+1]][j1]))
    }
```

```
}
return(c_eps)
}

#################################################################
Função 5 - Calcula os alfa_eps=d(m,h)*c_eps para cada sub-intervalo
em B, com d(m,h)=(1/h)4^m, h a variar entre 2^(-6) a 2^6
#################################################################
alfaeps<-function(B,C,h){
M <- length(B)
alfa_eps <- as.list(1:M)

for (m in 0:(M-1)){
    for(j in 1:2**m){
        j0 <- (j-1)*2 + 1
        j1 <- (j-1)*2 + 2
        alfa_eps[[m+1]][j0] <- (1/h)*(4**(m+1))*(1/sqrt(C[[m+1]][j0]))
        alfa_eps[[m+1]][j1] <- (1/h)*(4**(m+1))*sqrt(C[[m+1]][j0])
    }
}
return(alfa_eps)
}

#################################################################
Função 6 -   Calcula os logpsi(thetal)
#################################################################
logpsi_thetal<-function(B,x,A,B1,B2){
n <- length(x)
M <- length(B)
```

```
logpsi<-matrix(NA,(n-1),M)

for (i in 1:(n-1)){
    for (m in 0:(M-1)){
        aux <- B2[[m+1]][i+1]
        if (aux%%2==1) {
        # número de observações inferiores a xi à esquerda (B0)
            aux1 <- length(which(x[B1[[m+1]][[aux]]]<x[i+1]))
        # número de observações inferiores a xi à direita  (B1)
            aux2 <- length(which(x[B1[[m+1]][[aux+1]]]<x[i+1]))
            naux <- (A[[m+1]][aux]+aux1)*(A[[m+1]][aux]+A[[m+1]][aux+1])
            daux <- A[[m+1]][aux]*(A[[m+1]][aux]+aux1+A[[m+1]][aux+1]+aux2)
            logpsi[i,m+1] <- log(naux)-log(daux)
            # se pertencer ao intervalo da direita, B1
            } else {
            # número de observações inferiores a xi à esquerda (B0)
                aux1 <- length(which(x[B1[[m+1]][[aux]]]<x[i+1]))
            # número de observações inferiores a xi à direita  (B1)
                aux2 <- length(which(x[B1[[m+1]][[aux-1]]]<x[i+1]))
                naux <- (A[[m+1]][aux]+aux1)*(A[[m+1]][aux-1]+A[[m+1]][aux])
                daux <- A[[m+1]][aux]*(A[[m+1]][aux]+aux1+A[[m+1]][aux-1]+aux2)
                logpsi[i,m+1] <- log(naux)-log(daux)

        }
    }
}
return(logpsi)
}
```

Referências bibliográficas

Aitkin, M. (1991). Posterior Bayes factor (with discussion). *Journal of the Royal Statistal Society B*, *53*, 111-142.

Anderson, T. W., e Darling, D. A. (1954). A test of goodness-of-fit. *Journal of American Statistical Association*, *49*, 765-769.

Andrews, D. F., e Herzberg, A. M. (1985). *Data. A Collection of Problems from Many Fields for the Student and Research Worker.* New York: Springer-Verlag.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, *2*, 1152-1174.

Baringhaus, L., e Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Annals of the Institute of Statistical Mathematics*, *43*, 551-564.

Baringhaus, L., e Henze, N. (2000). Tests of fit for exponentiality based on a characterization via the mean residual life function. *Statistical Papers*, *41*, 225-236.

Barron, A., Schervish, M. J., e Wasserman, L. (1999). Posterior distributions in nonparametric problems. *The Annals of Statistics*, *27*, 536-561.

Bayarri, M. J., e Berger, J. O. (1997). Measures of Surprise in Bayesian Analysis. *ISDS Discussion Paper*, 97-46, Duke University.

Bayarri, M. J., e Berger, J. O. (1999). Quantifying surprise in the data and model verification. In J. M. Bernardo, J. O. Berger, A. P. Dawid, e A. F. M. Smith (Eds.), *Bayesian Statistics 6* (p. 53-82). London: Oxford University Press. Retrieved from `citeseer.ist.psu.edu/bayarri98quantifying.html`

Bayarri, M. J., e Berger, J. O. (2000). P-values for composite null models. *Journal of*

the American Statistical Association, *95*, 1127-1142.

Bayarri, M. J., e Castellanos, M. E. (2001). A comparison between p-values for goodness-of-fit checking. In *Monographs of Official Statistics. Bayesian methods with applications to science, policy and official statistics.* 1-10. Eurostat.

Berger, J. O., e Guglielmi, A. (2001). Bayesian Testing of a Parametric Model versus Nonparametric Alternatives. *Journal of the American Statistical Association*, *96*, 174-184.

Berger, J. O., e Pericchi, L. R. (1993). *The intrinsic Bayes factor for model selection* (Tech. Rep.). Department of Statistics, Purdue University, West Lafayette.

Berger, J. O., e Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and predition. *Journal of the American Statistical Association*, *91*, 109-122.

Blackwell, D., e MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, *1*, 353-355.

Box, G. E. P. (1980). Sampling and Bayes inferences in scientific modeling and robustness. *Journal of the Royal Statistical Society*, *Series A, 143*, 383-430.

Carlin, B. P., e Louis, T. A. (2000). *Bayesian and empirical Bayes methods for data analysis* (2nd ed.). New York: Chapman and Hall/CRC.

Carota, C., e Parmigiani, G. (1996). On Bayes Factors for Nonparametric Alternatives. In J. M. Bernardo, J. O. Berger, A. P. Dawid, e A. F. M. Smith (Eds.), *Bayesian Statistics 5* (p. 507-511). London: Oxford University Press.

Carota, C., Parmigiani, G., e Polson, N. G. (1996). Diagnostic measures for model criticism. *Journal of the American Statistical Association*, *91*, 753-762.

Chen, C. F. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society, Series B*, *47*, 540-546.

Chernoff, H., e Lehmann, E. L. (1954). The use of maximum likelihood estimates in chi-squared tests for goodness of fit. *The Annals of Mathematical Statistics*, *25*, 579-586.

Choi, B., Kim, K., e Song, S. H. (2004). Goodness of fit test for exponentiality based on Kullback-Leibler information. *Communications in Statistics - Simulation and*

Computation, *33(2)*, 525-536.

Coin, D. (2008). A goodness-of-fit test for normality based on polynomial regression. *Computational Statistics & Data Analysis*, *52*, 2185-2198.

Conigliani, C., Castro, J. I., e O'Hagan, A. (2000). Bayesian assessment of goodness-of-fit against nonparametric alternatives. *The Canadian Journal of Statistics*, *28*(2), 327-342.

Conigliani, C., e O'Hagan, A. (1996). *Sensitivity measures of the fractional Bayes factor* (Tech. Rep.). University of Nottingham.

Cox, D. R., e Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: Univ. Press.

D'Agostino, R. B., e Stephens, M. A. (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker Inc.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Epps, T. W., e Pulley, L. B. (1986). A test for exponentiality vs. monotone hazard alternatives derived from the empirical characteristic function. *Journal of the Royal Statistical Society, Series B*, *48(2)*, 206-213.

Escobar, M. D. (1988). *Estimating the means of several normal populations by nonparametric estimation of the distributions of the means* (Unpublished doctoral dissertation). Deparment of Statistics, Yale University.

Escobar, M. D. (1994). Estimating Normal Means with a Dirichlet Process Prior. *Journal of the American Statistical Association*, *89*, 268-277.

Escobar, M. D., e West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, *90*, 577-588.

Fabius, J. (1964). Asymptotic behavior of Bayes' estimates. *The Annals of Mathematical Statistics*, *35*, 846-856.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209-230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The*

Annals of Statistics, *2*, 615-629.

Florens, J. P., Richard, J. F., e Rolin, J. M. (1996). *Bayesian encompassing specification tests of a parametric model against a nonparametric alternative* (Tech. Rep.). Université Catholique de Louvain, Institut de Statistique.

Freedman, D. (1963). On the asymptotic distribution of Bayes' estimates in the discrete case. *Annals of Mathematical Statistics*, *34*, 1386-1403.

Gamerman, D., e Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (2nd ed.). Chapman and Hall/CRC.

Geisser, S., e Eddy, W. (1979). A predictive approach to model selection. *Journal of American Statistical Association*, *74*, 153-160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, e D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in Practice* (p. 145-161). London: Chapman and Hall.

Gelfand, A. E., Dey, D., e Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, e A. F. M. Smith (Eds.), *Bayesian Statistics 4* (p. 147-167). London: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., e Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., e Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). London: Chapman and Hall/CRC.

Gelman, A., e Meng, X.-L. (1996). Model checking and model improvement. In W. R. Gilks, S. Richardson, e D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in Practice* (p. 189-202). London: Chapman and Hall.

Gelman, A., Meng, X.-L., e Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*, 733-807.

Ghosh, J. K., e Ramamoorthi, R. V. (2003). *Bayesian Nonparametric*. New York: Springer-Verlag.

Grané, A., e Fortiana, J. (2011). A directional test of exponentiality based on maximum

correlations. *Metrika*, *73*, 255-274.

Guttman, I. (1967). The use of concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society*, *B*, 83-100.

Hanson, T. (2006). Inference for Mixtures of Finite Polya Tree Models. *Journal of the American Statistical Association*, *101*, 1548-1565.

Hanson, T., e Johnson, W. O. (2002). Modeling regression errors with a mixture of Polya trees. *Journal of the American Statistical Association*, *97*, 1020-1033.

Henze, N., e Meintanis, S. G. (2005). Recent and classical tests for exponentiality: A partial review with comparisons. *Metrika*, *61*, 29-45.

Hjort, N. L., Dahl, F. A., e Steinbakk, G. H. (2006). Post-processing posterior predictive p-values. *Journal of the American Statistical Association*, *101*, 1157-1174.

Hjort, N. L., Holmes, C., Müller, P., e Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Jara, A. (2008). *Bayesian semiparametric methods for the analysis of complex data* (Unpublished doctoral dissertation). University Leuven.

Jeffreys, H. (1961). *The Theory of Probability* (3rd ed.). London: Oxford University Press.

Johnson, V. E. (2004). A Bayesian chi-squared test for goodness-of-fit. *The Annals of Statistics*, *32:6*, 2361-2384.

Kass, R. E., e Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*, 773-795.

Koehler, K. J., e Gan, F. F. (1990). Chi-squared goodness-of-fit tests: Cell selection and power. *Communications in Statistics - Simulation and Computation*, *19*, 1265-1278.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *The Annals of Statistics*, *20*, 1222-1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *The Annals of Statistics*, *22*, 1161-1176.

Mann, H. B., e Wald, A. (1942). On the choice of the number of class intervals in

the application of the chi-square test. *The Annals of Mathematical Statistics*, *13*, 306-317.

Mauldin, R. D., Sudderth, W. D., e Williams, S. C. (1992). Polya trees and random distributions. *The Annals of Statistics*, *20*, 1203-1221.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*, 1142-1160.

O'Hagan, A. (1991). Discussion of posterior Bayes factor (by M. Aitkin). *Journal of the Royal Statistical Society, Series B*, *53*, 136.

O'Hagan, A. (1995). Fractional Bayes factor for model comparison (with discussion). *Journal of the Royal Statistical Society*, *B, 56:1*, 99-138.

O'Hagan, A. (1997). Properties of intrinsic and fractional Bayes factor. *Test*, *6*, 101-118.

Paddock, S. M. (1999). *Randomized Polya Trees: Bayesian Nonparametrics for Multivariate Data Analysis* (Unpublished doctoral dissertation). Institute of Statistics and Decision Sciences, Duke University.

Paddock, S. M., Ruggeri, F., Lavine, M., e West, M. (2003). Randomized Polya tree models for nonparametric Bayesian inference. *Statistica Sinica*, *13*, 443-460.

Paulino, C. D., Amaral Turkman, M. A., e Murteira, B. (2003). *Estatística bayesiana.* Fundação Calouste Gulbenkian, Lisboa.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 157-175.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Robins, J. M., van der Vaart, A., e Ventura, V. (2000). Asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association*, *95*, 1143-1159.

Rolin, J. (1992). *Some useful properties of the Dirichlet Process* (Tech. Rep.). Université

Catholique de Louvain, Belgium.

Romão, X., Delgado, R., e Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation*, *80:5*, 545-591.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*, 1151-1172.

Schervish, M. J. (1995). *Theory of Statistics.* New York: Springer.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639-650.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., e van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583-639.

Thode, C. H. (2002). *Testing for Normality.* New York: Marcel Dekker.

Tokdar, S. T., e Martin, R. (2011). *Bayesian test of normality versus a Dirichlet process mixture alternative* (Tech. Rep.). Manuscript Submitted for Publication, 2011.

Verdinelli, I., e Wasserman, L. A. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *The Annals of Statistics*, *20*, 1203-1221.

Wakefield, J., e Walker, S. (1997). Bayesian Nonparametric Population Models: Formulation and Comparison with Likelihood Approaches. *Jounal of Pharmacokinetics and Biopharmaceuties*, *25:2*, 235-253.

Walker, S. G., e Mallick, B. K. (1997). Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing. *Journal of the Royal Statistical Society*, *Series B, 59*, 845-860.

Zhang, J., e Wu, Y. (2005). Likelihood-ratio tests for normality. *Computational Statistics and Data Analysis*, *49*, 709-721.